

# Partitioned Sampling of Public Opinions Based on Their Social Dynamics

Weiran Huang<sup>\*</sup>  
IIIS<sup>†</sup>, Tsinghua University  
Beijing, China  
huang.inbox@outlook.com

Liang Li  
Microsoft Research Asia  
Beijing, China  
liangl@microsoft.com

Wei Chen  
Microsoft Research Asia  
Beijing, China  
weic@microsoft.com

## ABSTRACT

Public opinion polling is typically done by random sampling from the entire population, treating the opinions of individuals as independent. In the real world, individuals' opinions are often correlated, especially among friends in a social network, due to the effect of both homophily and social influence. In this paper, we explore the idea of partitioned sampling, which partition individuals likely holding similar opinions into groups and sample every group separately to obtain an accurate estimate of the population opinion. We first rigorously formulate the above idea into an optimization problem. In particular, we characterize individuals' opinions as random variables, specify the objective as minimizing the expected sample variance of the estimated result, and precisely define the statistical measure of pairwise opinion similarity, which by our analysis is enough to fully determine the solution of the optimization problem. We show that the simple partitions that contain only one sample in each group are always better, and reduce finding the optimal simple partition to a well-studied Max- $r$ -Cut problem. We adopt the semi-definite programming algorithm for Max- $r$ -Cut to solve our optimization problem, and further develop a greedy heuristic to improve efficiency. Moreover, to address the issue of how to obtain opinion similarity efficiently, we propose an opinion evolution model based on reasonable user interaction patterns in social network, and provide efficient and exact computation of opinion similarity in the model. We use both synthetic and real-world datasets to demonstrate that our partitioned sampling method results in significant improvement in sampling quality.

## CCS Concepts

•**Theory of computation** → *Sketching and sampling; Random walks and Markov chains*; •**Human-centered computing** → *Social networks*;

## Keywords

sampling; social networks; opinion evolution dynamics

## 1. INTRODUCTION

<sup>†</sup>Institute for Interdisciplinary Information Sciences

<sup>\*</sup>This work was done when the first author was visiting Microsoft Research Asia as a research intern. This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003, 61433014.

Public opinion polling is a main tool for governments, organizations and companies to gather information about public sentiments on the policies, strategies, products etc., which are important in organizational decision making. Opinion polling needs to be accurate and unbiased, and thus it is usually done by randomly sampling a large enough number of individuals from the entire population, but this is a costly effort. Therefore, saving the cost on unbiased random sampling while keeping the same sampling quality is an important task to pursue.

In this paper, we utilize individuals' social interactions to improve the random sampling method. Our motivation is that people's opinions are often correlated, especially among friends in the social network, due to their social interactions in terms of the homophily and influence effects [5, 11, 18]. In the era of big data, these social interactions and correlations are partially known. For example, many online social media and social networking sites provide public available social interaction data and users' sentiment data, and companies also have large amounts of data about their customers' preferences and their social interactions. Our idea is to partition individuals into different groups by utilizing these partial knowledge, such that people within the same group are likely to hold similar opinions on a topic of interest. We can then sample a few people in each group and aggregate the samplings together to achieve an accurate sampling result. We call this *partitioned sampling* method.

We formulate the above idea as an optimization problem. In particular, we first characterize individuals' opinions as random variables, which are likely to be correlated due to their social interactions. We then specify our objective as minimizing the expected sample variance of the estimated result, and define the statistical measure of pairwise *opinion similarity* as the input. Our analysis later show that this input is enough to fully determine the solution of the optimization problem, which we call as the *Optimal Partitioned Sampling (OPS)* problem (Section 2).

We solve the OPS problem and reach two important conclusions (Section 3). First, we show that the best partition is always the *simple partition*, meaning that each group only contains one sample. Second, we use people's opinion similarities to construct a weighted graph and reduce the OPS problem to a graph partitioning problem, which is a special case of the well-studied Max- $r$ -Cut problem. We then adopt the semi-definite programming algorithm for Max- $r$ -Cut to solve the OPS problem, and further propose an efficient greedy partitioning algorithm to work on larger graphs.

To address the issue of how to efficiently obtain pairwise opinion similarities, we propose an opinion evolution model, *Voter model with Innate Opinions (VIO)*, which is adapted from the voter model often used to characterize opinion evolution dynamics in social networks (Section 4). We provide efficient and exact computation of opinion similarity in the steady state of the model, which can be used as the input to the OPS problem.

Finally, we conduct experiments on both synthetic and real-world datasets to demonstrate that our partitioned sampling method indeed improves sampling quality over traditional naive sampling method, which translates into significant cost savings if we maintain sampling quality at the same level (Section 5).

In summary, our contributions include: (a) proposing the partitioned sampling method and formulating it as an optimization problem to improve sampling quality based on people’s opinion similarities; (b) precisely connecting the OPS problem to the Max- $r$ -Cut problem and providing efficient algorithms for the OPS problem, and (c) adapting an opinion evolution model and providing an exact and efficient method for computing opinion similarities.

Due to space constraint, we put most of the mathematical proofs and technical details into the appendix.

**Related work.** To the best of our knowledge, there is no other technical work on partitioned sampling. Among studies on population sampling, Dasgupta et al. [7] also utilize social network connections to facilitate sampling. However, their method is to explicitly ask the subject being sampled to return additional information about their friends’ opinions and the number of their friend’s friends, which requires changing the polling practice. Our partitioned sampling method, on the other hand, still follows the standard polling practice and only uses implicit knowledge on opinion correlations to improve sampling quality. These two ideas are orthogonal and could be potentially combined together. Another popular social-interaction based sampling method is respondent-driven sampling [14, 15], which gets individuals to refer those they know. It is an adaptive sampling method and designed for estimating the proportion of those very small and hard-to-reach groups such as drug injectors, which are hard to deal with by normal sampling methods. Our partitioned sampling method is not limited by the size of the target population. Das et al. [6] study the task of removing the correlations among individual’s opinions due to their social interactions to obtain the average *original innate opinion*. Their task is to utilize the wisdom of the crowd for extracting the latent *independent* opinions of individuals. Our task is exactly the opposite — we want to utilize opinion interactions and correlations for more efficient sampling of *final expressed opinions*, which are what being counted for in opinion polling.

Various opinion evolution models have been proposed in the literature (e.g., [6, 10, 16, 21]). Our VIO model and its analysis are adapted from the voter model [2] and its extension with stubborn agents [21]. The models in [6, 10] also distinguish between innate opinions and expressed opinions, however, their models are deterministic, and thus their analyses do not apply to our stochastic analysis on the similarities and correlations between final expressed opinions.

Graph partitioning has been well studied, and numerous problem variants and algorithms exist. In this paper, we reduce the OPS problem to the Max- $r$ -Cut problem, which

---

### Method 1 Naive Sampling

---

**Require:** Vertex set  $V$ , sample size  $r$ .

- 1: Choose  $r$  sample nodes  $x_1, x_2, \dots, x_r$  from  $V$  uniformly at random with replacement.
  - 2: **Output:**  $\hat{f}_{naive}(V, r) = \frac{1}{r} \sum_{i=1}^r f(x_i)$ .
- 

---

### Method 2 Partitioned Sampling

---

**Require:** Partition  $\mathcal{P} = \{(V_1, r_1), (V_2, r_2), \dots, (V_K, r_K)\}$ .

- 1: **for**  $k \leftarrow 1$  **to**  $K$  **do**
  - 2:   Do naive sampling in  $V_k$ , **return**  $\hat{f}_{naive}(V_k, r_k)$ .
  - 3: **end for**
  - 4: **Output:**  $\hat{f}_{part}(\mathcal{P}) = \sum_{k=1}^K \frac{|V_k|}{|V|} \cdot \hat{f}_{naive}(V_k, r_k)$ .
- 

is the problem of partitioning the graph into  $r$  groups and maximizing the sum of edge weights on the cut, and we adopt the semidefinite programming algorithm proposed in [8] for solving our OPS problem.

## 2. FORMULATING THE PARTITIONED SAMPLING PROBLEM

We consider a vertex set  $V$  from a social network graph containing  $n$  vertices (or nodes)  $v_1, v_2, \dots, v_n$ . Each vertex represents a person in the social network, and has a binary opinion on some topic of interest. Our task is to do efficient sampling with sample size budget  $r$  for estimating the average opinion of all the people in the social network. Let  $f : V \rightarrow \{0, 1\}$  denote the opinion function, i.e., we wish to estimate the fraction  $\bar{f} = \frac{1}{n} \sum_{i=1}^n f(v_i)$ . The most basic sampling method is called *naive sampling*, which picks  $r$  people uniformly at random with replacement from  $V$  to ask their opinions and takes the average of sampled opinions as the estimate (Method 1). We denote the estimate of naive sampling as  $\hat{f}_{naive}(V, r) = \frac{1}{r} \sum_{i=1}^r f(x_i)$ , where  $x_i$  is the  $i$ -th sampled person.

In this paper, we propose a general sampling framework called *partitioned sampling* (Method 2). Formally, we first partition the whole vertex set into several disjoint subsets (called *groups*) such that each vertex is included in one and only one of the groups, and then allocate subsample size of each group. We use notation  $\mathcal{P} = \{(V_1, r_1), (V_2, r_2), \dots, (V_K, r_K)\}$  to represent such a partition, where  $V_1, V_2, \dots, V_K$  are disjoint groups with  $\bigcup_{k=1}^K V_k = V$ , and  $r_k$  is the subsample size of group  $V_k$  with  $\sum_{k=1}^K r_k = r$ . Next, we do naive sampling inside each group  $V_k$  with its corresponding subsample size  $r_k$ . Finally, we estimate the average opinion of the entire population by taking a weighted average of all subsampling results, where the weight is proportional to the size of the group. We use  $\hat{f}_{part}(\mathcal{P})$  to denote the final estimate of partitioned sampling using partition  $\mathcal{P}$ . It is worth mentioning that naive sampling is a special case of partitioned sampling, when all vertices in  $V$  are partitioned into the unique group  $\mathcal{P} = \{(V, r)\}$ . Before using partitioned sampling framework to design the efficient sampling method, we first show its unbiasedness in the following proposition.

**PROPOSITION 1. (Unbiasedness)** *Partitioned sampling is unbiased. Specifically, for any partition  $\mathcal{P}$ ,  $\mathbb{E}[\hat{f}_{part}(\mathcal{P})] = \bar{f}$ .*

**PROOF.** For any partition  $\mathcal{P} = \{(V_1, r_1), \dots, (V_K, r_K)\}$ ,

according to the output formulation of Method 2,

$$\begin{aligned}\mathbb{E}[\hat{f}_{part}(\mathcal{P})] &= \sum_{k=1}^K \frac{|V_k|}{|V|} \mathbb{E}[\hat{f}_{naive}(V_k, r_k)] \\ &= \frac{1}{|V|} \sum_{k=1}^K |V_k| \sum_{v_i \in V_k} \frac{f(v_i)}{|V_k|} = \frac{\sum_{i=1}^n f(v_i)}{|V|} = \bar{f}.\end{aligned}$$

Notice that naive sampling in any group  $V_k$  is unbiased, thus  $\mathbb{E}[\hat{f}_{naive}(V_k, r_k)]$  is equal to the average opinion of the people in  $V_k$  (second equality above). Therefore partitioned sampling is unbiased.  $\square$

Intuitively, the advantage of using partitioned sampling is that, if we partition individuals such that people likely holding the same opinions are partitioned into the same group, then we can sample very few people in each group to get an accurate estimate of the average opinion of the group, and aggregate them to get a good estimate of population opinion. To implement this idea, we need to have some prior knowledge about people’s opinions before doing sampling, such as which people will be likely to hold the same opinions and how much likelihood they have. Such prior knowledge can be derived from the past history of people’s expressed opinions, or their social interaction dynamics. Based on this knowledge, our goal is to find the best partition for partitioned sampling which achieves the best sampling efficiency.

Our first research challenge is how to rigorously formulate the above intuition into an optimization problem. To meet this challenge, we need to answer (a) which objective function is the appropriate one for the optimization problem, and (b) which representation of the prior knowledge about people’s opinions and their similarities can be used as the input to the optimization problem to enable solving the problem.

We first address the objective function. When individuals’ opinions  $f(v_1), f(v_2), \dots, f(v_n)$  are fixed (but unknown), by convention, the effectiveness of an unbiased randomized sampling method is measured by the sample variance  $\text{Var}(\hat{f})$ , where  $\hat{f}$  is the estimated result based on the randomized sampling method. The smaller the sample variance, the better the sampling method. When partial statistical knowledge about people’s opinions are available, effectively we treat opinions  $f(v_1), f(v_2), \dots, f(v_n)$  as random variables, and the partial knowledge is some statistics related to the joint distribution of these random variables. In this case, the best sampling method should minimize the *expected sample variance*  $\mathbb{E}[\text{Var}(\hat{f})]$ , where the expectation is taken over the randomness from the joint distribution of opinions. To clarify the source of randomness, henceforth we use subscript  $M$  (standing for “model”) to represent the randomness from joint distribution model of the opinions, and subscript  $S$  (standing for “sampling”) to represent sample randomness from the sampling method, and thus  $\mathbb{E}[\text{Var}(\hat{f})]$  is clarified as  $\mathbb{E}_M[\text{Var}_S(\hat{f})]$ . One may also propose to use the total variance  $\text{Var}_{M,S}(\hat{f})$  combining randomness from both the joint distribution and the sampling method. In fact, we show in Appendix B, that these two objective functions are equivalent. Therefore, our objective function is settled as minimizing the expected variance  $\mathbb{E}_M[\text{Var}_S(\hat{f})]$ .

We now answer the question of what kind of statistical knowledge about the joint distribution of people’s opinions  $f(v_1), f(v_2), \dots, f(v_n)$  is needed as the input to enable the

minimization of  $\mathbb{E}_M[\text{Var}_S(\hat{f})]$ . First, note that these random variables are not mutually independent. In fact, people’s opinions are often correlated due to their social interactions, and partitioned sampling is based on the very fact about the existence of these correlations and would be useless if all people’s opinions are independent. Thus, to fully characterize the joint distribution, we may need an exponential number of parameters, which is definitely infeasible in practice. Our first attempt is to use the expectations and the correlations of these random variables (corresponding to the first two moments of the joint distribution) as the input for the optimization problem. Indeed we find that these information is good enough to fully characterize the optimization problem we need to solve. However, by a deeper investigation, we discover that a weaker and more direct type of statistics is equivalent in terms of fully characterizing the optimization problem, which we formally define as pairwise opinion similarities: For binary random variables  $f(v_1), f(v_2), \dots, f(v_n)$ , the *opinion similarity*  $\sigma_{ij}$  for node pair  $v_i$  and  $v_j$  is defined as the probability that  $f(v_i)$  and  $f(v_j)$  have the same values. Intuitively, the above defined similarity would help us to partition the nodes likely to hold the same opinions together. In fact, in the next section, we will show our key result that knowing pairwise similarities is enough for us to reduce the optimization problem to a graph partitioning problem (Theorem 1).

With the objective function and problem inputs settled, we are now ready to define our optimization problem:

**DEFINITION 1.** (*Optimal Partitioned Sampling*) Given a vertex set  $V = \{v_1, v_2, \dots, v_n\}$  and sample size budget  $r < n$ . Suppose the opinion similarity  $\sigma_{ij}$  between any pair of nodes  $v_i$  and  $v_j$  is known before sampling. The Optimal Partitioned Sampling (OPS) problem is to find the optimal partition  $\mathcal{P}^*$  of  $V$  with the corresponding sample size allocation, such that the partitioned sampling method as given in Method 2 with the above partition  $\mathcal{P}^*$  achieves the minimum expected sample variance, i.e.,  $\mathcal{P}^* = \arg \min_{\mathcal{P}} \mathbb{E}_M[\text{Var}_S(\hat{f}_{part}(\mathcal{P}))]$ , where  $\mathcal{P}$  takes among all partitions of  $V$  with  $r$  samples.

### 3. SOLVING THE OPS PROBLEM

There are two issues involved in the OPS problem: One is how to partition the vertex set  $V$  into  $r$  groups; The other is how to allocate the subsample size of each group. For simplifying the OPS problem, we first consider a special kind of partition that picks only one sample node in each group.

**DEFINITION 2.** A *simple partition* is the partition such that the subsample size of each group is equal to one.

Simple partition is important not only for its simplicity but also for its superiority. We will later show in Theorem 2 that for any non-simple partition, there always exists a better simple partition, which can also be efficiently constructed. Henceforth, we focus on finding the optimal simple partition for the OPS problem.

Our approach is to construct a weighted assistant graph  $G_a$  whose vertex set is  $V$ , where the weight of edge  $(v_i, v_j)$  is  $w_{ij} = 1 - \sigma_{ij}$ , and then connect the OPS problem with a graph partitioning problem for the graph  $G_a$ . For any simple partition  $\mathcal{P} = \{(V_1, 1), (V_2, 1), \dots, (V_r, 1)\}$  of  $V$ , we use  $\text{Vol}_{G_a}(V_k)$  to denote the volume of the group  $V_k$  in the graph

$G_a$ , given by  $\text{Vol}_{G_a}(V_k) = \sum_{v_i, v_j \in V_k} w_{ij}$ . We define a cost function  $g(\mathcal{P})$  to be the summation of all groups' volumes in  $G_a$ , namely,  $g(\mathcal{P}) = \sum_{k=1}^r \text{Vol}_{G_a}(V_k)$ . Our major technical contribution is to show that minimizing the expected sample variance of partitioned sampling using simple partitions of  $V$  is equivalent to minimizing the volume summation of all the groups in  $G_a$ , as summarized by the following theorem:

**THEOREM 1.** *Given a vertex set  $V = \{v_1, v_2, \dots, v_n\}$  with their pairwise opinion similarities  $\{\sigma_{ij}\}$ 's and sample size  $r < n$ , we construct an assistant graph  $G_a$  whose vertex set is  $V$  and edge  $(v_i, v_j)$ 's weight  $w_{ij}$  is  $1 - \sigma_{ij}$ . For any simple partition  $\mathcal{P} = \{(V_1, 1), (V_2, 1), \dots, (V_r, 1)\}$  of  $V$ ,*

$$\mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}))] = \frac{1}{2n^2} \cdot g(\mathcal{P})$$

where  $g(\mathcal{P})$  is the volume summation of all the groups of  $\mathcal{P}$  in  $G_a$ . Thus, the optimal simple partition of  $V$  minimizes the cost function  $g(\mathcal{P})$ .

The proof of the theorem requires involved computation and is moved to Appendix A due to space constraint. Theorem 1 provides the connection between the OPS problem and the graph partitioning problem. Intuitively, small cost function indicates small volume of each group, thus the edge weights between the nodes in the same group are small, which means the nodes in the same group have high similarities. Therefore, for the partitions with small cost function, the nodes in the same group tend to hold the same opinions, and thus the nodes being sampled can effectively represent the opinions of their respective groups. To sum up, Theorem 1 makes precise our intuition that grouping people with similar opinion tendencies together would make partitioned sampling more efficient.

Theorem 1 also suggests that we can reduce the OPS problem to the following *Min- $r$ -Volume* problem: Given an undirected graph with non-negative edge weights, partition the graph into  $r$  groups such that the summation of all groups' volumes is minimized. However, the Min- $r$ -Volume problem contains  $r$ -Coloring problem as a special case, which has minimum volume summation of zero if and only if the graph is  $r$ -colorable. This leads to the following strong inapproximability result:

**LEMMA 1.** *The Min- $r$ -Volume problem is NP-hard to be approximated within any finite factor.*

Note that the above hardness does not directly imply the hardness of the OPS problem, since the assistant graph  $G_a$  is a special form with particular edge weights which are generated by opinion similarities. We leave the hardness of the OPS problem as an open question, and next use the dual problem of Min- $r$ -Volume to help solving the OPS problem.

The dual problem of Min- $r$ -Volume is the following *Max- $r$ -Cut* problem: Given an undirected graph with non-negative edge weights, partition the graph into  $r$  groups such that the total edge weight of the cut (set of edges crossing different groups) is maximized. It is clear that Min- $r$ -Volume and Max- $r$ -Cut are equivalent in terms of exact solutions, but they are different in terms of approximability. In particular, Frieze and Jerrum [8] show that for the Max- $r$ -Cut problem, a semi-definite programming (SDP) based partition achieves  $1 - 1/r + 2 \ln r / r^2$  approximation ratio. We adopt the SDP algorithm to solve the OPS problem. The SDP algorithm

---

#### Algorithm 1 SDP Partitioning Algorithm

---

**Require:** Graph  $G_a$  with  $n$  nodes, partition size  $r$ .

- 1: Solve the following SDP problem and compute the Cholesky decomposition of  $Y$ . Let  $y_1, y_2, \dots, y_n$  be the resulting vectors.

$$\text{Maximize} \quad \frac{r-1}{r} \sum_{i \neq j} w_{ij} (1 - Y_{ij}) \quad (\text{SDP})$$

$$\text{Subject to} \quad \begin{aligned} & \text{(a) } Y_{ii} = 1, \forall i; \text{ (b) } Y_{ij} \geq -\frac{1}{r-1}, \forall i \neq j; \\ & \text{(c) } Y \succeq 0; \text{ (d) } Y \text{ is symmetric.} \end{aligned}$$

- 2: Choose  $r$  random vectors  $z_1, z_2, \dots, z_r$  from  $\mathbb{R}^n$ .
  - 3: Partition  $V$  into  $r$  groups  $V_1, \dots, V_r$  according to which of  $z_1, z_2, \dots, z_r$  is closest to each  $y_k$ .<sup>1</sup>
  - 4: **Output:**  $\mathcal{P} = \{(V_1, 1), \dots, (V_r, 1)\}$ .
- 

---

#### Algorithm 2 Greedy Partitioning Algorithm

---

**Require:** Graph  $G_a$  with  $n$  nodes, partition size  $r$ .

- 1: Randomly generate a node sequence of all the nodes:  $x_1, x_2, \dots, x_n$ .
  - 2: Let  $V_1 = \dots = V_r = \emptyset$ .
  - 3: **repeat**
  - 4:   **for**  $i \leftarrow 1$  **to**  $n$  **do**
  - 5:     **if**  $x_i \in V_j$  for some  $j \in [r]$  **then**  $V_j = V_j \setminus \{x_i\}$ .
  - 6:     **end if**
  - 7:      $k \leftarrow \arg \min_{\ell \in [r]} \delta g_\ell(x_i, \{(V_1, 1), \dots, (V_r, 1)\})$ .
  - 8:      $V_k \leftarrow V_k \cup \{x_i\}$ .
  - 9:   **end for**
  - 10: **until** a predetermined stopping condition holds.
  - 11: **Output:**  $\mathcal{P} = \{(V_1, 1), \dots, (V_r, 1)\}$ .
- 

including the SDP relaxation program is given in Algorithm 1 (the original integer programming formulation is given in Appendix C).

The drawback of the SDP partitioning algorithm is that it is rather slow. Thus we further propose a heuristic greedy algorithm to solve the Min- $r$ -Volume problem, which can be applied to larger graphs. Given a simple partition  $\mathcal{P} = \{(V_1, 1), \dots, (V_r, 1)\}$  of  $V$  and an external node  $v_i$  which is not in  $V$ , we define  $\delta g_\ell(v_i, \mathcal{P})$  to be  $g(\mathcal{P}') - g(\mathcal{P})$ , where  $\mathcal{P}'$  is equal to  $\{(V_1, 1), \dots, (V_\ell \cup \{v_i\}, 1), \dots, (V_r, 1)\}$ . Thus  $\delta g_\ell(v_i, \mathcal{P})$  represents the increase of the cost function when the external node  $v_i$  is added to the group  $V_\ell$  of  $\mathcal{P}$ .

The basic idea of our greedy algorithm (Algorithm 2) is to assign each ungrouped node  $x_i$  to a group such that the objective function  $g(\mathcal{P})$  is increased the least. After the first round of greedy assignment, we repeat the greedy assignment procedure to further decrease the cost function, until some stopping condition holds, such as the decrease is smaller than a predetermined threshold. Notice that the cost function is positive, and there exists a certain gap<sup>2</sup> that the decrease in each round must be greater than, therefore, the cost function should no longer decrease after finite rounds.

The running time of the one-round greedy partitioning is  $O(n + m)$  where  $m$  is the number of edges in the assistant

<sup>1</sup>If the partitioning result is less than  $r$  groups, just reselect  $r$  new random vectors from  $\mathbb{R}^n$  and repeat the step again.

<sup>2</sup>The gap is  $\min_{i, U, W} |\sum_{j \in U} w_{ij} - \sum_{k \in W} w_{ik}|$  where  $U$  and  $W$  are two subsets of  $V$  satisfying  $U \cup W \subseteq V \setminus \{v_i\}$  and  $U \cap W = \emptyset$ .



graph  $G_a$ . In our experiment, we will show that greedy partitioning with a reasonable stopping condition performs as well as SDP partitioning but could run on much larger graphs.

The performance of partitioned sampling using the simple partition generated by the greedy partitioning algorithm is always at least as good as naive sampling, even using the partition generated only after the first round of greedy assignment, as summarized below:

LEMMA 2. *Given a vertex set  $V$  with  $n$  nodes and sample size  $r < n$ , partitioned sampling using the simple partition  $\mathcal{P}$  generated by the greedy partitioning algorithm (even after the first round) is at least as good as naive sampling. Specifically,*

$$\mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}))] \leq \mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{naive}}(V, r))].$$

We call a partition  $\mathcal{P}'$  a *refined* partition of  $\mathcal{P}$ , if each group of  $\mathcal{P}'$  is a subset of some group of  $\mathcal{P}$ . Suppose we are given a partition  $\mathcal{P}$  such that there exists some groups containing at least two sample nodes. Then we can further partition each of those groups by the greedy partitioning algorithm and finally obtain a refined simple partition of  $\mathcal{P}$ . According to Lemma 2, the refined simple partition would be at least as good as the original partition  $\mathcal{P}$ . This leads us to the following theorem.

THEOREM 2. *For any non-simple partition  $\mathcal{P}$ , there exists a refined simple partition  $\mathcal{P}'$  of  $\mathcal{P}$ , which can be constructed efficiently, such that partitioned sampling using the refined simple partition  $\mathcal{P}'$  is at least as good as partitioned sampling using the original partition  $\mathcal{P}$ . Specifically,*

$$\mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}'))] \leq \mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}))].$$

This theorem shows the superiority of simple partitions, and justifies that it is enough for us to only optimize for partitioned sampling with simple partitions.

## 4. OPINION EVOLUTION MODEL

So far, we have proposed the partitioned sampling method and pointed out the way to obtain the optimal partition for the OPS problem. However, to apply the partitioned sampling method to practice, there exists one more issue of how to get people's opinion similarities before doing sampling, which are taken as inputs in the OPS problem. The most straightforward way to obtain any two people's opinion similarity is gathering enough historical voting data which they attended both or their sentiment data for some similar topics, and then computing the proportion of the frequency that they expressed the same opinions. In fact, people's opinions are often interrelated in the real world, which can be characterized by the opinion evolution models. Thus, a more efficient way to address the issue is to derive people's opinion similarities from the opinion evolution models. In this section, we first propose a new opinion evolution model called VIO model in Section 4.1. Based on this model, we show the way for obtaining people's opinion similarities in Section 4.2. At last, we provide an efficient computation of opinion similarities in Section 4.3.

### 4.1 VIO Model Description

In social networks, one's opinion is often affected by her friends, leading to opinion clustering in the network. Voter

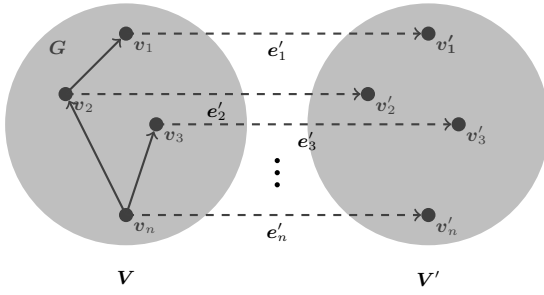
model [2] is a popular one used to describe such opinion dynamics, and various extensions exist, such as the model in [21] that allows stubborn agents who do not change their own opinions. We further extend the model in [21] to allow a person to either keep her own innate opinion or adopt a friend's opinion (similar in concepts as the model in [6, 10]), and also allow different individuals to update their opinions with distinct rates. More specifically, each node in the social graph is associated with both an *innate* opinion and an *expressed* opinion for any given topic. The innate opinion remains unchanged from external influences, while the expressed opinion could be shaped by the opinions of one's neighbors, and is the one observed by sampling. We call this adapted model *Voter model with Innate Opinions (VIO)*, and describe its technical detail below.

We consider a weighted directed social graph  $G = (V, A)$  where  $V = \{v_1, v_2, \dots, v_n\}$  is the vertex set containing  $n$  vertices and  $A$  is the weighted adjacency matrix. An edge  $(v_i, v_j)$  exists if the entry  $A_{ij} > 0$ . Each node represents an individual in the social network, and edge  $(v_i, v_j)$  represents the opinion influence relationship from  $v_j$  to  $v_i$ . For each node  $v_i$  in the graph, let  $f^{(t)}(v_i) \in \{0, 1\}$  denote its expressed opinion at time  $t$ , for  $t \geq 0$ . At initial time  $t = 0$ , each node  $v_i$  generates its innate opinion  $f^{(0)}(v_i)$  from an i.i.d. Bernoulli distribution with expected value  $\mu^{(0)}$ . The use of i.i.d. distribution for the initial opinion is due to the lack of prior knowledge when the initial opinion for a brand-new topic is formed, and has been used in other models before (e.g., [7]). Each node  $v_i$  updates its expressed opinion according to its Poisson process with rate  $\lambda_i$  independently, and keeps its expressed opinion unchanged at other times. Each time, there is only one node updating its expressed opinion. In particular, node  $v_i$ , at its Poisson arrival time  $t$ , sets its expressed opinion  $f^{(t)}(v_i)$  to be its own innate opinion  $f^{(0)}(v_i)$  with an *inward probability*  $p_i$ , or with probability  $1 - p_i$ , node  $v_i$  randomly selects one of its out-neighbors  $v_j$  with probability proportional to the weight of the edge  $(v_i, v_j)$  (i.e., with probability  $(1 - p_i)A_{ij} / \sum_{k=1}^n A_{ik}$ ) and sets its expressed opinion  $f^{(t)}(v_i)$  to be  $f^{(t)}(v_j)$ . Our model degenerates to the original voter model when all the inward probabilities are equal to zero and all the Poisson rates are identical, and to the model with stubborn agents [21] when all stubborn agents have  $p_i = 1$  and other agents have  $p_i = 0$ , meanwhile, all agents have  $\lambda_i = 1$ . Thus the inward probability  $p_i$  represents the inward tendency (or stubbornness) of node  $v_i$ . In summary, our VIO model is parametrized by the weighted adjacency matrix  $A$ , the inward probabilities  $p_1, p_2, \dots, p_n$ , the Poisson rates  $\lambda_1, \lambda_2, \dots, \lambda_n$ , and the expected value of innate opinion  $\mu^{(0)}$ .

The VIO model reaches a steady state if the joint distribution of all people's expressed opinions no longer changes over time. Our first observation is that the VIO model has a unique steady state, as summarized as follow:

LEMMA 3. *When  $p_i > 0$  for all  $i \in [n]$ , the VIO model has a unique joint distribution for the final expressed opinions in the steady state.*

We use notation  $f^{(\infty)}(v_i)$  to represent the steady-state expressed opinion of node  $v_i$ , which is a random variable. We assume that opinion sampling is done after the system reaches the steady state. This means that people have sufficiently communicated within the social network about



**Figure 1: The augmented graph  $\bar{G}$ :** The left gray circle represents the original social graph  $G = (V, A)$ ; The right gray circle represents the added vertex set  $V'$ , the copy of  $V$ . The dashed lines represent the new directed edges connecting the corresponding nodes between  $V$  and  $V'$ .

their opinions on the topic of interest before doing sampling, which is a reasonable assumption.

## 4.2 Random-Walk Based Analysis

We now provide the analysis of the VIO model, the key quantities we want to obtain from the evolution model are the similarities of people's expressed opinions in the steady state. In fact, opinion similarities can be derived from opinion expectations and opinion correlations, thus our approach is first studying people's expressed opinion expectations and opinion correlations, and then obtaining the opinion similarities from them. We use notation  $\mu_i$  and  $\rho_{ij}$  to represent the expectation of  $v_i$ 's expressed opinion and the correlation between  $v_i$ 's and  $v_j$ 's expressed opinions in the steady state, respectively. From now on, we only study the VIO model with  $p_i > 0$  for all  $i \in [n]$ , that is, individuals always leave some chance for their innate opinions.

To facilitate the analysis, we first construct an augmented graph  $\bar{G}$  (Figure 1) from the original social graph  $G$  as follows (similar to the construction in [10]). Based on the social graph  $G = (V, A)$ , we add a new vertex set  $V' = \{v'_1, v'_2, \dots, v'_n\}$ , which is a copy of the vertex set  $V$ . Each vertex  $v_i \in V$  connects to its corresponding vertex  $v'_i \in V'$  with a directed edge  $e'_i = (v_i, v'_i)$ . Thus the augmented graph  $\bar{G} = (V \cup V', E \cup \{e'_1, e'_2, \dots, e'_n\})$  is established where  $E$  is the edge set of  $G$ .

The voter model and its variants are often analyzed through the equivalent *coalescing random walks* (e.g., [4, 17, 21]). We now specify the coalescing random walk model on the augmented graph  $\bar{G}$ , which is equivalent to the VIO model. In order to track the influence of nodes' expressed opinions and finally obtain nodes' expressed opinions at time  $t$ , we consider  $n$  walkers starting random walks on the graph  $\bar{G}$  from time  $t$ , but "back in time": At time  $t$ , the  $n$  random walkers are located at nodes  $v_1, v_2, \dots, v_n$ , respectively. Suppose  $v_i$  is the last node who updated its expressed opinion before time  $t$  and the updating happened at time  $\tau < t$ . Then the  $n$  walkers stay on the nodes from time  $t$  until time  $\tau$  "back in time", and at time  $\tau$ , the walker who is at node  $v_i$  take a walk step. She either walks to  $v_i$ 's out-neighbor  $v_j \in V$  with probability  $(1 - p_i)A_{ij} / \sum_{k=1}^n A_{ik}$ , or walks to  $v'_i \in V'$  with probability  $p_i$ . This random-walk step is exactly like the step of node  $v_i$  when it decides which opinion to adopt at time  $\tau$  in the VIO model. It represents that  $v_i$ 's

expressed opinion at time  $t$  comes from the node where the walker walks to at time  $\tau$ . After this random walk step, all the walkers stay put waiting for the next walk step, which is taken at the last Poisson arrival before time  $\tau$  "forward in time". When the Poisson arrives, the walker who is located at the corresponding node at that time<sup>3</sup>, takes a step in the same manner. At any time, if the walker who starts at node  $v_i$  reaches a node  $v'_k \in V'$ , then she stops her walk (is absorbed), and  $v_i$ 's opinion at time  $t$  is determined to be  $v_k$ 's innate opinion, that is  $f^{(t)}(v_i) = f^{(0)}(v_k)$ . If two random walkers meet at the same node in  $V$  at any time, then they will walk together from now on following the above rule (hence the name *coalescing*). Finally, at time  $t = 0$ , if the walker is still at some node  $v_i \in V$ , she always walks to  $v'_i \in V'$ .

It is straightforward to verify that the coalescing random walk model is equivalent to the VIO model, in that for every fixed innate opinions  $f^{(0)}(v_1), f^{(0)}(v_2), \dots, f^{(0)}(v_n)$ , the joint distribution of  $f^{(t)}(v_1), f^{(t)}(v_2), \dots, f^{(t)}(v_n)$  of the VIO model is the same as the joint distribution of the nodes' innate opinions where the  $n$  walkers finally stand when they all reach  $V'$ , if they start their walks at nodes  $v_1, v_2, \dots, v_n$  respectively at time  $t$ .

Note that when we study the steady state of the VIO model, the time  $t$  tends to infinity, and since  $p_i > 0$  for all  $i$ , all random walkers reach  $V'$  before time 0 with probability 1, thus the special random walk rule for  $t = 0$  is not essential. Thus, we also say that the steady state behavior is when all random walkers start their random walks at time  $t = \infty$ . Since the steady state is unique and reachable as shown in Lemma 3, thus we can analyze the stochastic quantities of the steady state. First, we show that the expectation of any node  $v_i$ 's final expressed opinion is equal to the expected value of innate opinion.

**LEMMA 4.** *The expected expressed opinion of each node in the steady state is equal to the expected value of innate opinion, namely, for all  $i \in [n]$ ,*

$$\mu_i = \mathbb{E}_M[f^{(\infty)}(v_i)] = \mu^{(0)}.$$

We then focus on studying the opinion correlations between each pair of nodes. To do so, we provide some key definitions related to the coalescing random walk model, together with their analysis below.

**DEFINITION 3.** *Let  $\mathcal{I}_{ij}^\ell$  denote the event that two random walks starting from  $v_i$  and  $v_j$  at time  $t = \infty$  eventually meet and the first node they meet at is  $v_\ell \in V$ . Let  $Q$  be the  $n \times n$  matrix where  $Q_{ij}$  denotes the probability that a random walker starting from node  $v_i$  at time  $t = \infty$  ends at  $v'_j \in V'$ .*

The following lemma provides the exact computation of parameter  $\mathbb{P}[\mathcal{I}_{ij}^\ell]$  and  $Q$ .

<sup>3</sup>If there is no walker located at the corresponding node at that time, then no walk step happens and all walkers stay put for the next Poisson arrival. Such a situation occurs in the case that two consecutive Poisson arrivals come at the same node, and in this case, the first Poisson arrival (which is the second Poisson arrival "back in time") is noneffective and can be ignored.

LEMMA 5. For  $i, j, \ell \in [n]$ ,  $\mathbb{P}[\mathcal{I}_{ij}^\ell]$  is the unique solution of the following linear equation system:

$$\mathbb{P}[\mathcal{I}_{ij}^\ell] = \begin{cases} 0, & i = j \neq \ell, \\ 1, & i = j = \ell, \\ \sum_{a=1}^n \frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i+\lambda_j)d_i} \mathbb{P}[\mathcal{I}_{aj}^\ell] \\ + \sum_{b=1}^n \frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i+\lambda_j)d_j} \mathbb{P}[\mathcal{I}_{ib}^\ell], & i \neq j, \end{cases}$$

where  $d_i = \sum_{j=1}^n A_{ij}$  is node  $v_i$ 's weighted out-degree. In addition, matrix  $Q$  is computed by

$$Q = (I - (I - P)D^{-1}A)^{-1}P,$$

where  $P = \text{diag}(p_1, p_2, \dots, p_n)$  and  $D = \text{diag}(d_1, d_2, \dots, d_n)$  are two diagonal matrices, and matrix  $I - (I - P)D^{-1}A$  is invertible when  $p_i > 0$  for all  $i \in [n]$ .

With  $\{\mathbb{P}[\mathcal{I}_{ij}^\ell]\}$ 's and  $Q$  computed, we can obtain the correlation between any two expressed opinions in the steady state. The following lemma provides our main analytical result concerning the VIO model.

LEMMA 6. For any  $i, j \in [n]$ , opinion correlation  $\rho_{ij}$  in the steady state is equal to the probability that two coalescing random walks starting from  $v_i$  and  $v_j$  at time  $t = \infty$  end at the same absorbing node in  $V'$ . Moreover, opinion correlation  $\rho_{ij}$  can be computed by

$$\begin{aligned} \rho_{ij} &= \text{Cor}_M(f^{(\infty)}(v_i), f^{(\infty)}(v_j)) \\ &= \sum_{k=1}^n Q_{ik} Q_{jk} + \sum_{\ell=1}^n \mathbb{P}[\mathcal{I}_{ij}^\ell] \left(1 - \sum_{k=1}^n Q_{\ell k}^2\right) \end{aligned}$$

where  $\mathcal{I}_{ij}^\ell$  and  $Q$  are defined in Definition 3, and  $\mathbb{P}[\mathcal{I}_{ij}^\ell]$  and  $Q$  are computed by Lemma 5.

With opinion expectations  $\{\mu_i\}$ 's and opinion correlations  $\{\rho_{ij}\}$ 's calculated by Lemma 4 and Lemma 6, we now turn to compute the opinion similarities  $\{\sigma_{ij}\}$ 's in the steady state, which are requested in the OPS problem as inputs.

THEOREM 3. For any two nodes  $v_i$  and  $v_j$ , their opinion similarity  $\sigma_{ij}$  in the steady state of the VIO model is equal to:

$$\sigma_{ij} = 1 - 2\mu^{(0)}(1 - \mu^{(0)})(1 - \rho_{ij})$$

where opinion correlation  $\rho_{ij}$  is computed by Lemma 6.

Recall that in Theorem 1, we reduce the OPS problem into the Min- $r$ -Volume problem for the assistant graph  $G_a$ , whose edge weight is  $1 - \sigma_{ij}$ . Theorem 3 indicates that  $1 - \sigma_{ij}$  equals to  $1 - \rho_{ij}$  times a constant  $2\mu^{(0)}(1 - \mu^{(0)})$ . Therefore, for finding the optimal partition for the OPS problem with the VIO model, which minimizes the volume summation of  $G_a$ , the exact value of  $\mu^{(0)}$  is needless and what we only need to compute are the opinion correlations  $\{\rho_{ij}\}$ 's. In the following section, we will provide an efficient way to calculate them.

### 4.3 Efficient Correlation Computation

Naive correlation computation directly using Lemma 5 and 6 by solving the linear equation system for  $\{\mathbb{P}[\mathcal{I}_{ij}^\ell]\}$ 's would have a running time of  $O(n^7)$  (See in proof of Lemma 5). We now improve the running time to  $O(nmR)$

by a carefully designed iterative computation method, where  $m$  is the number of edges of the social graph  $G$  and  $R$  is the number of iterations.

We use  $q_{ij}^{(k)}$  to represent the probability that a random walker starting at  $v_i$  is absorbed at  $v_j' \in V'$  after  $k$  walk steps in the coalescing random walk model. Thus  $q_{ij}^{(0)} = 0$  and  $q_{ij}^{(k)}$  can be computed iteratively by:<sup>4</sup>

$$q_{ij}^{(k)} = p_i \cdot 1_{i=j} + \sum_{a=1}^n \frac{(1-p_i)A_{ia}}{d_i} q_{aj}^{(k-1)}.$$

The running time of computing  $\{q_{ij}^{(k)}\}$ 's with one iteration is

$$\sum_{i=1}^n \sum_{j=1}^n (1 + d_i) = O(nm),$$

where  $m$  is number of edges of the social graph  $G$ .

Now we consider two walkers take coalescing random walks on the graph  $\overline{G}$  starting at  $v_i$  and  $v_j$  respectively. We use notation  $M_{ij}^{(k)}$  to denote the probability that their walks coalesce after they taking altogether  $k$  walk steps. Notice that  $M_{ij}^{(0)} = 0$  for  $i \neq j$ , and  $M_{ij}^{(0)} = 1$  for  $i = j$ . For  $k \geq 1$ ,  $M_{ij}^{(k)}$  can be calculated iteratively by:<sup>4</sup>

$$\begin{aligned} M_{ij}^{(k)} &= \frac{\lambda_i}{\lambda_i + \lambda_j} \left[ p_i q_{ji}^{(k-1)} + \sum_{a=1}^n \frac{(1-p_i)A_{ia}}{d_i} M_{aj}^{(k-1)} \right] \\ &+ \frac{\lambda_j}{\lambda_i + \lambda_j} \left[ p_j q_{ij}^{(k-1)} + \sum_{a=1}^n \frac{(1-p_j)A_{ja}}{d_j} M_{ai}^{(k-1)} \right]. \end{aligned}$$

The running time of computing  $\{M_{ij}^{(k)}\}$ 's with one iteration is

$$\sum_{i=1}^n \sum_{j=1}^n (1 + d_i) + (1 + d_j) = O(nm),$$

where  $m$  is number of edges of the social graph  $G$ .

According to Lemma 6, opinion correlation  $\rho_{ij}$  in the steady state is equal to  $\lim_{k \rightarrow \infty} M_{ij}^{(k)}$ . Thus we can obtain opinion correlations by computing  $\{M_{ij}^{(k)}\}$ 's and  $\{q_{ij}^{(k)}\}$ 's iteratively in time  $O(nmR)$  until reaching some convergence precision, where  $R$  is the number of iterations.

## 5. EXPERIMENTAL EVALUATION

In this section, we present results of our experimental evaluations of the sampling quality of our partitioning algorithms proposed in Section 3 compared against naive sampling based on the VIO model, using both synthetic and real-world datasets. To be specific, we first describe the generation of our synthetic graphs, and show the comparison of various sampling methods and how graph structure and inward tendency affect the performance of the sampling methods in Section 5.1. We then move on to the real-world dataset in Section 5.2 to show a method of learning the distribution of inward probabilities from online social networks, and the performance of partitioned sampling on the real-world graph based on the learned inward probabilities.

In our experiment, when the parameters of VIO model (i.e., weighted adjacency matrix  $A$ , people's inward probabilities  $p_1, p_2, \dots, p_n$ , updating rates of people's opinions

<sup>4</sup>The proof and analysis can be found in Appendix D.

$\lambda_1, \lambda_2, \dots, \lambda_n$ , and the expected value of innate opinion  $\mu^{(0)}$  are set, the experiment are done by (a) calculating the similarities of every pair of nodes by the method given in Theorem 3 and Section 4.3, (b) running the partitioning algorithms<sup>5</sup> to obtain the partition candidate, and (c) computing the expected variance  $\mathbb{E}_M[\text{Var}_S(\hat{f})]$ <sup>6</sup> by Theorem 1. In both synthetic and real-world datasets, we set  $\mu^{(0)}$  to 0.5. Notice that the value of  $\mu^{(0)}$  has no effect on the results (Theorem 3).

## 5.1 Synthetic Dataset

In our synthetic experiments, we use the hidden partition model [3] to generate the undirected graphs, which aims at resembling the community structure in real-world social networks. It is specified by four parameters: the number of vertices  $n$ , the number of hidden partitions  $k$ , the inter-partition and intra-partition edge probabilities  $p_H$  and  $p_L$ , respectively. First, we assign each node to one of the  $k$  hidden partitions uniformly at random. Next, we independently connect each pair of nodes in the same hidden partition with probability  $p_H$ , and two nodes in different partitions with probability  $p_L < p_H$ . We generate two different sizes of hidden partition graphs. The small one is used to compare the sampling quality among the following three different sampling methods: naive sampling (Naive), partitioned sampling using greedy partitioning (Greedy), and partitioned sampling using SDP partitioning (SDP). We use a small graph because SDP is infeasible to run on large graphs. We then move on to the large hidden partition graphs and only run Greedy and Naive for those graphs. We study the impact of inward probabilities and graph structure on the performance of Greedy on those graphs, respectively. For the synthetic graphs, we set opinion updating rate  $\lambda_i$  to 1 for all  $i \in [n]$ .

**Small synthetic graph.** The small hidden partition graphs we generate includes 100 nodes and 20 hidden partitions. Probability  $p_H$  and  $p_L$  are set to 0.9 and 0.01, respectively. The inward probability of each node is randomly chosen from  $[0, 0.01]$ . We range the sample size  $r$  from 2 to 14, and the expected sample variance is shown in Figure 2(a). We put sample size on  $y$ -axis to make it easier to see the savings on the sample size under the same expected sample variance.

When the sample size  $r$  is small (i.e., less than 8), the performance of SDP and Greedy are similar to each other, and both better than Naive. When the sample size  $r$  increases, the performance of Greedy becomes much better than Naive, and the performance of SDP starts getting worse but it is still better than Naive. Specifically, if we fix  $\mathbb{E}_M[\text{Var}_S(\hat{f})]$  to be 0.01, Greedy and SDP need 8 samples, while Naive needs 12 samples. This suggests that by using our partitioned sampling method, we can save 33% of samples while achieving the same sampling quality.

<sup>5</sup>To solve the SDP programming in SDP partitioning algorithm, we used CVX which is a package for specifying and solving convex programs [12, 13].

<sup>6</sup>Each randomized partitioning algorithm was run 10 times, and we took the average of the expected variance as the result. In the OPS problem, for the objective function, the outermost randomness is the randomness  $M$ , but in the experiment as above, the outermost randomness is the randomness from randomized algorithms, which should be included in the randomness  $S$ . We show in Appendix B that they are equivalent.

**Large synthetic graphs.** For the large synthetic graph with 10k nodes and 500 hidden partitions, SDP is no longer feasible, thus we compare the improvement of Greedy against Naive. We run Greedy and Naive using different sample sizes ( $r = 250$  and  $r = 500$ ), varying the inward probabilities and  $p_H/p_L$ , to observe the improvement of expected sample variance under different graph clustering and inward tendency settings.

In Figure 2(b), we set all nodes' inward probabilities to 0.05 and  $p_H$  to 1, and range  $p_L$  from  $10^{-1}$  to  $10^{-6}$ . The improvement on  $y$ -axis means the improvement of expected sample variance from Naive to Greedy. The result shows that the larger  $p_H/p_L$  (more apparent clustering) indicates the better performance of our partitioned sampling. When  $p_H/p_L$  increases from  $10^2$  to  $10^5$ , the improvement of expected sample variance enhances rapidly. When  $p_H/p_L$  is too large (i.e., larger than  $10^5$ ), the improvement of expected sample variance becomes saturated. This is because the number of edges which cross different hidden partitions are very few so that it decreases rather slowly and the graph structure is almost unchanged when  $p_H/p_L$  increases further.

In Figure 2(c), we set  $p_H$  to 1 and  $p_L$  to be  $10^{-5}$  to generate the hidden partition graph. For this graph, we set all nodes' inward probabilities to be identical, varying from 0.02 to 0.8. The result shows that the lower inward probability indicates the better performance of our partitioned sampling. When the inward probability is small, the improvement of expected sample variance increases rapidly. This is because a lower inward probability means people interact more with their friends and thus their opinions are correlated more significantly.

According to the above two experiments, we conclude that the larger  $p_H/p_L$  and the lower inward probability make people's opinions clustered and strongly correlated inside the clusters, and our partitioned sampling method works better for these cases.

## 5.2 Real-World Dataset

The real-network dataset we use is the online social network data from Sina Weibo<sup>7</sup> [22], which contains 100,102 users and 30,518,600 tweets within a one-year timeline from 1/1/2013 to 1/1/2014. We treat the user following relationship between two users as a directed edge (with weight 1). For this dataset, we first need to learn the distribution of people's inward probabilities.

### 5.2.1 Distribution of Inward Probabilities

In order to observe the evolution of opinions for a specific topic of interest, We manually choose 12 specific topics (e.g., Microsoft, iPhone, etc.), and extract all tweets from the Weibo dataset related to these topics (simply using keyword based classifier). We then run each tweet through a sentiment analyzer [20] to obtain binary opinion values (positive/negative). Thus we get a series of opinions for each user at discrete time corresponding to each topic. For each topic, we select those users who published opinions at least 4 times, and regard their first opinions as their innate opinions  $f^{(0)}(v_1), f^{(0)}(v_2), \dots, f^{(0)}(v_n)$  and treat the average of the rest opinions as their expected opinions  $\mu_1, \mu_2, \dots, \mu_n$  in the steady state state. We then collect their relationships and form a subgraph for the corresponding topic.

<sup>7</sup><http://weibo.com>



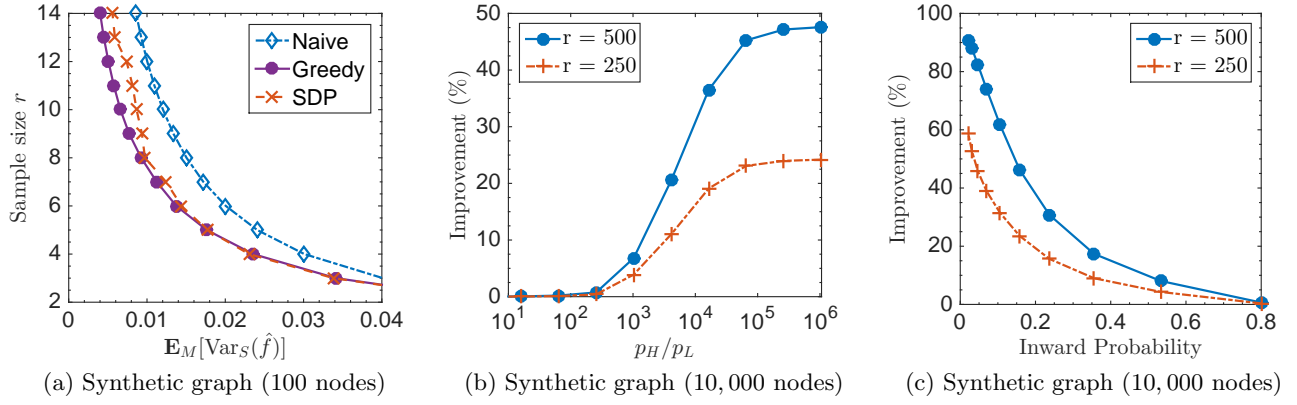


Figure 2: Experimental results on synthetic graphs.

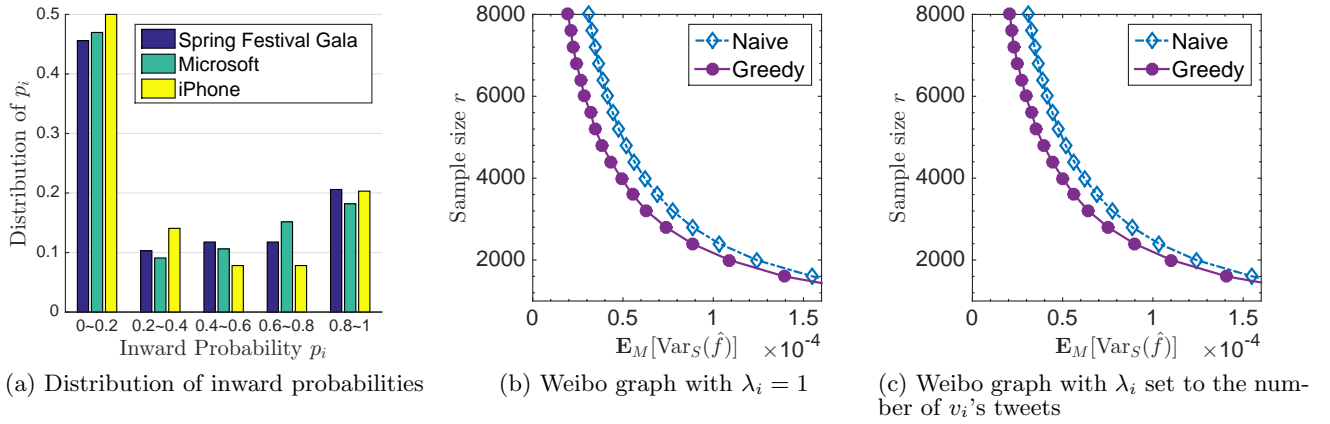


Figure 3: Experimental results on real-world graphs.

Recall the definition of matrix  $Q$  (Definition 3), and it is easy to see that

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = Q \begin{pmatrix} f^{(0)}(v_1) \\ f^{(0)}(v_2) \\ \vdots \\ f^{(0)}(v_n) \end{pmatrix} \quad (\text{or } \vec{\mu} = Q\vec{f}^{(0)}).$$

Thus we can estimate the inward probabilities by solving the following programming

$$\begin{aligned} & \text{Minimize } \|\vec{\mu} - Q\vec{f}^{(0)}\|, \\ & \text{Subject to } 0 \leq p_i \leq 1, \forall i \in [n], \end{aligned}$$

and we use gradient descent method to handle above programming.

We estimate the inward probabilities under the 12 topics respectively, and Figure 3(a) shows the distribution of inward probabilities for three topics, namely Spring Festival Gala (68 users), Microsoft (66 users) and iPhone (59 users), and the results for other topics are similar. The distribution for these three different topics are quite similar: (a) Over 45% inward probabilities locate in  $[0, 0.2]$ ; (b) The probability that  $p_i$  locates in  $[0.8, 1]$  is the second highest; (c) Others almost uniformly locate in  $[0.2, 0.8]$ . This indicates that in the real world, most people tend to adopt others' opinions,

which matches the intuition that many people are affected by other people's opinions. We manually look up those users who locate in  $[0.8, 1]$ , and find that a large number of them are media accounts and verified users. This matches our intuition that those users always take effort to spread their own opinions on the web but less likely to adopt other people's opinions, hence they should have large inward probabilities.

### 5.2.2 Performance of Partitioned Sampling

In this section, we show the performance of Greedy compared to Naive on the real-world graph. For the original Weibo dataset, we first remove the users who do not follow anyone, iteratively. Then we get our Weibo graph including 40,787 nodes and 165,956 directed edges. We use two different settings for opinion updating rates: One is to set  $\lambda_i = 1$  for all  $i \in [n]$ ; The other is to set  $\lambda_i$  to be the number of  $v_i$ 's tweets from 1/1/2013 to 1/1/2014 in the Weibo dataset. The users' inward probabilities are set in the following way so that it follows the distribution we learned in the previous section: We sort all the inward probabilities learned in Section 5.2.1 among 12 topics, denoted as  $\hat{p}_1 \leq \hat{p}_2 \leq \dots \leq \hat{p}_k$ . For each user  $v_i$  in the Weibo graph, we select an integer  $j$  from  $\{1, 2, \dots, k+1\}$  uniformly at random, and set  $v_i$ 's inward probability to a random real number in the following

interval

$$\begin{cases} [0, \hat{p}_1], & \text{if } j = 1, \\ [\hat{p}_k, 1], & \text{if } j = k + 1, \\ [\hat{p}_{j-1}, \hat{p}_j], & \text{others.} \end{cases}$$

Since there are some  $\hat{p}_j$  values that are zeros, we will have users with zero inward probability. For these users, we use a very small value  $10^{-10}$  in our simulation since our computation of the VIO model requires inward probability to be greater than zero.

Figure 3(b) and 3(c) show the experimental results on the Weibo graph with all  $\lambda_i = 1$  and  $\lambda_i$  set to the number of  $v_i$ 's tweets, respectively. The improvement of Greedy against Naive with two different updating rate settings are similar. In particular, if we fix  $\mathbb{E}_M[\text{Var}_S(\hat{f})]$  to be  $3.101 \times 10^{-5}$ , Greedy needs 5715 samples while Naive needs 8000 samples (saving 28.6%) in Figure 3(b), and Greedy needs 5811 samples while Naive needs 8000 samples (saving 27.4%) in Figure 3(c).

The results indicate that the sample size saving is more apparent when the expected sample variance is getting smaller (i.e., sampling quality requirement is higher). The figures also indicate that our partitioned sampling method is robustly better than naive sampling method regardless of the updating rate settings. The results are consistent with the results from synthetic graphs in demonstrating the better performance of partitioned sampling method.

In summary, our results on the real-world data show that real-world social networks do exhibit opinion correlations and clusterings that can enable more efficient sampling through the partitioned sampling method. Our results on the synthetic data further show that when graph clustering and social interaction are stronger, the benefit of partitioned sampling could be higher.

## 6. DISCUSSION AND FUTURE WORK

For the OPS problem, if the inputs learned from the data are not accurate, our partitioned sampling method using either SDP or greedy approach may lose its effectiveness. However, if we make a small modification of the partitioning algorithms to force the output partition to be a balanced partition (having the equal size for each group), the balanced partitioning algorithms always achieve better sampling quality than naive sampling, no matter whether the opinion similarity estimation is accurate or not. We show the proof in Appendix E due to the space constraint.

There are a number of open problems and future directions one may pursue. For example, one may further enrich the VIO model to allow (a) non-identical innate opinion distributions if partial knowledge about individuals' innate opinion tendencies is available, or (b) negative relationships as modeled in [16] so as to include negative correlations, and study the opinion similarities under these models. We provide the results on the above two issues in Appendix F and G. Another direction is to improve learning the parameters of VIO model, or in general to extract opinion similarities in social networks from real-world data. We remark that our results on the OPS problem is not constraint to the VIO model, and thus other efficient methods for opinion similarity estimation, if available, can also be applied together with the OPS optimization algorithms.

## References

- [1] S. Brakken-Thal. Gershgorin's theorem for estimating eigenvalues, 2007.
- [2] P. Clifford and A. Sudbury. A model for spatial conflict. *Biometrika*, 1973.
- [3] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures & Algorithms*, 2001.
- [4] J. T. Cox. Coalescing random walks and voter model consensus times on the torus in  $\mathbb{Z}^d$ . *The Annals of Probability*, 1989.
- [5] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. KDD '08, 2008.
- [6] A. Das, S. Gollapudi, R. Panigrahy, and M. Salek. Debiasing social wisdom. KDD '13, 2013.
- [7] A. Dasgupta, R. Kumar, and D. Sivakumar. Social sampling. KDD '12, 2012.
- [8] A. Frieze and M. Jerrum. Improved approximation algorithms for max k-cut and max bisection. *Algorithmica*, 1997.
- [9] M. R. Garey and D. S. Johnson. Computer and intractability. *A Guide to the Theory of NP-Completeness*, 1979.
- [10] A. Gionis, E. Terzi, and P. Tsaparas. Opinion maximization in social networks. SDM '13, 2013.
- [11] S. Goel, W. Mason, and D. J. Watts. Real and perceived attitude agreement in social networks. *Journal of Personality and Social Psychology*, 2010.
- [12] M. Grant and S. Boyd. Graph implementations for non-smooth convex programs. In *Recent Advances in Learning and Control*. 2008.
- [13] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming. <http://cvxr.com/cvx>, 2014.
- [14] D. D. Heckathorn. Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, 1997.
- [15] D. D. Heckathorn. Respondent-driven sampling ii: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 2002.
- [16] Y. Li, W. Chen, Y. Wang, and Z. Zhang. Voter model on signed social networks. *Internet Mathematics*, 2015.
- [17] T. M. Liggett. *Interacting particle systems*. Springer Science & Business Media, 2006.
- [18] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 2001.
- [19] J. R. Norris. *Markov chains*. Cambridge university press, 1998.
- [20] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. ACL '14, 2014.
- [21] E. Yildiz, D. Acemoglu, A. E. Ozdaglar, A. Saberi, and A. Scaglione. Discrete opinion dynamics with stubborn agents. *Available at SSRN 1744113*, 2011.
- [22] N. J. Yuan, F. Zhang, D. Lian, K. Zheng, S. Yu, and X. Xie. We know how you live: Exploring the spectrum of urban lifestyles. COSN '13, 2013.

## APPENDIX

### A. MATHEMATICAL PROOFS

**THEOREM 1.** *Given a vertex set  $V = \{v_1, v_2, \dots, v_n\}$  with their pairwise opinion similarities  $\{\sigma_{ij}\}$ 's and sample size  $r < n$ , we construct an assistant graph  $G_a$  whose vertex set is  $V$  and edge  $(v_i, v_j)$ 's weight  $w_{ij}$  is  $1 - \sigma_{ij}$ . For any simple partition  $\mathcal{P} = \{(V_1, 1), (V_2, 1), \dots, (V_r, 1)\}$  of  $V$ ,*

$$\mathbb{E}_M[\text{Var}_S(\hat{f}_{part}(\mathcal{P}))] = \frac{1}{2n^2} \cdot g(\mathcal{P})$$

where  $g(\mathcal{P})$  is the volume summation of all the groups of  $\mathcal{P}$  in  $G_a$ . Thus, the optimal simple partition of  $V$  minimizes the cost function  $g(\mathcal{P})$ .

**PROOF.** We use  $x_k$  to denote the sample node selected in the  $k$ -th group  $V_k$  of the simple partition  $\mathcal{P}$ .

The estimate of partitioned sampling using partition  $\mathcal{P}$  can be written as

$$\hat{f}_{part}(\mathcal{P}) = \frac{\sum_{k=1}^r n_k f(x_k)}{n}$$

where  $n_k$  is the size of  $k$ -th group  $V_k$ . Thus the sample variance of  $\hat{f}_{part}(\mathcal{P})$  is equal to

$$\begin{aligned} \text{Var}_S[\hat{f}_{part}(\mathcal{P})] &= \mathbb{E}_S[\hat{f}_{part}(\mathcal{P})^2] - \mathbb{E}_S[\hat{f}_{part}(\mathcal{P})]^2 \\ &= \mathbb{E}_S\left[\left(\frac{\sum_{k=1}^r n_k f(x_k)}{n}\right)^2\right] - \bar{f}^2 \\ &= \mathbb{E}_S\left[\frac{1}{n^2} \sum_{k=1}^r n_k^2 f(x_k) + \frac{2}{n^2} \sum_{k < l} n_k n_l f(x_k) f(x_l)\right] - \bar{f}^2 \\ &= \frac{1}{n^2} \sum_{k=1}^r n_k^2 \mathbb{E}_S[f(x_k)] + \frac{2}{n^2} \sum_{k < l} n_k n_l \mathbb{E}_S[f(x_k)] \mathbb{E}_S[f(x_l)] - \bar{f}^2 \\ &= \frac{1}{n^2} \sum_{k=1}^r n_k^2 \bar{f}_k + \frac{2}{n^2} \sum_{k < l} n_k n_l \bar{f}_k \bar{f}_l - \bar{f}^2, \end{aligned}$$

where  $\bar{f}$  is the average opinion of the entire population and  $\bar{f}_k$  is the average opinion of the  $k$ -th group.

We take the expectation of each item in the above equation under the randomness  $M$ :

$$\begin{aligned} a) \quad \mathbb{E}_M\left[\frac{1}{n^2} \sum_{k=1}^r n_k^2 \bar{f}_k\right] &= \frac{1}{n^2} \sum_{k=1}^r n_k \mathbb{E}_M[n_k \bar{f}_k] \\ &= \frac{1}{n^2} \sum_{k=1}^r n_k \sum_{v_i \in V_k} \mathbb{E}_M[f(v_i)], \end{aligned}$$

$$\begin{aligned} b) \quad \mathbb{E}_M\left[\frac{2}{n^2} \sum_{k < l} n_k n_l \bar{f}_k \bar{f}_l\right] &= \frac{2}{n^2} \sum_{k < l} \mathbb{E}_M\left[\left(\sum_{v_i \in V_k} f(v_i)\right) \left(\sum_{v_j \in V_l} f(v_j)\right)\right] \\ &= \frac{2}{n^2} \sum_{k < l} \sum_{v_i \in V_k} \sum_{v_j \in V_l} \mathbb{E}_M[f(v_i) f(v_j)], \end{aligned}$$

$$\begin{aligned} c) \quad \mathbb{E}_M[\bar{f}^2] &= \mathbb{E}_M\left[\left(\frac{\sum_{i=1}^n f(v_i)}{n}\right)^2\right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_M[f(v_i)^2] + \frac{2}{n^2} \sum_{i < j} \mathbb{E}_M[f(v_i) f(v_j)] \\ &= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_M[f(v_i)] + \frac{2}{n^2} \sum_{k=1}^r \sum_{\substack{i < j \\ v_i, v_j \in V_k}} \mathbb{E}_M[f(v_i) f(v_j)] \\ &\quad + \frac{2}{n^2} \sum_{k < l} \sum_{v_i \in V_k} \sum_{v_j \in V_l} \mathbb{E}_M[f(v_i) f(v_j)]. \end{aligned}$$

Thus

$$\begin{aligned} \mathbb{E}_M[\text{Var}_S(\hat{f}_{part}(\mathcal{P}))] &= \mathbb{E}_M\left[\frac{1}{n^2} \sum_{k=1}^r n_k^2 \bar{f}_k\right] + \mathbb{E}_M\left[\frac{2}{n^2} \sum_{k < l} n_k n_l \bar{f}_k \bar{f}_l\right] - \mathbb{E}_M[\bar{f}^2] \\ &= \frac{1}{n^2} \sum_{k=1}^r n_k \sum_{v_i \in V_k} \mathbb{E}_M[f(v_i)] \\ &\quad - \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_M[f(v_i)] - \frac{2}{n^2} \sum_{k=1}^r \sum_{\substack{i < j \\ v_i, v_j \in V_k}} \mathbb{E}_M[f(v_i) f(v_j)] \\ &= \frac{2}{n^2} \sum_{k=1}^r \left[ \frac{n_k - 1}{2} \sum_{v_i \in V_k} \mathbb{E}_M[f(v_i)] - \sum_{\substack{i < j \\ v_i, v_j \in V_k}} \mathbb{E}_M[f(v_i) f(v_j)] \right]. \end{aligned}$$

Notice that for any two binary (0/1) random variables  $A$  and  $B$ , the following equation holds:

$$\mathbb{E}[AB] = \frac{1}{2} (\mathbb{P}[A = B] + \mathbb{E}[A] + \mathbb{E}[B] - 1).$$

Thus

$$\mathbb{E}_M[f(v_i) f(v_j)] = \frac{1}{2} (\sigma_{ij} + \mathbb{E}_M[f(v_i)] + \mathbb{E}_M[f(v_j)] - 1) \quad (1)$$

Therefore,

$$\begin{aligned} \mathbb{E}_M[\text{Var}_S(\hat{f}_{part}(\mathcal{P}))] &= \frac{1}{n^2} \sum_{k=1}^r \sum_{\substack{i < j \\ v_i, v_j \in V_k}} (1 - \sigma_{ij}) \\ &= \frac{1}{2n^2} \cdot \sum_{k=1}^r \text{Vol}_{G_a}(V_k) = \frac{1}{2n^2} \cdot g(\mathcal{P}). \end{aligned}$$

Ignoring constant terms, the optimal simple partition  $\mathcal{P}$  minimizes the cost function  $g(\mathcal{P})$ .  $\square$

**LEMMA 1.** *The Min- $r$ -Volume problem is NP-hard to be approximated within any finite factor.*

**PROOF.** We establish a reduction from the  $r$ -coloring problem to the Min- $r$ -Volume problem as follows.

An  $r$ -coloring of the graph is an assignment of one of  $r$  possible colors to each vertex such that the endpoints of every edge are colored differently. The  $r$ -coloring problem is to decide if there exists an  $r$ -coloring for a given graph. For  $r > 2$ , this problem is NP-complete [9]. Suppose we have an approximation algorithm for Min- $r$ -Volume on the graph  $G$  with finite approximation factor. We can use this

algorithm to solve an instance of  $r$ -coloring of the graph  $G$  in polynomial time, in the following manner.

If the optimal solution of Min- $r$ -Volume for the graph  $G$  is zero, there should be no edges inside any group. Then we color the vertices in the same group with the same color, thus there will be no two adjacent vertices sharing the same color. This is a  $r$ -coloring of the graph  $G$ . Otherwise, if there exists a  $r$ -coloring of the graph  $G$ , we put the same colored vertices in the same group, leading to the sum of  $r$  volumes equal to zero. If we apply the approximation algorithm of Min- $r$ -Volume with finite approximation factor to the graph  $G$ , we are able to distinguish whether the optimal solution of Min- $r$ -Volume is zero or not, which indicates whether the  $r$ -coloring of the graph  $G$  exists or not.

Hence, we establish the polynomial-time reduction.  $\square$

LEMMA 2. *Given a vertex set  $V$  with  $n$  nodes and sample size  $r < n$ , partitioned sampling using the simple partition  $\mathcal{P}$  generated by the greedy partitioning algorithm (even after the first round) is at least as good as naive sampling. Specifically,*

$$\mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}))] \leq \mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{naive}}(V, r))].$$

PROOF. It has been known that, the sample variance of naive sampling is

$$\text{Var}_S(\hat{f}_{\text{naive}}(V, r)) = \frac{1}{r}(\bar{f} - \bar{f}^2)$$

where  $\bar{f}$  is the average opinion of the entire population and  $\mathbb{E}_M[\bar{f}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_M[f(v_i)]$ . According to the Eq. (1),

$$\mathbb{E}_M[f(v_i)f(v_j)] = \frac{1}{2}(\sigma_{ij} + \mathbb{E}_M[f(v_i)] + \mathbb{E}_M[f(v_j)] - 1).$$

Thus

$$\begin{aligned} \mathbb{E}_M[\bar{f}^2] &= \mathbb{E}_M\left[\left(\frac{\sum_{i=1}^n f(v_i)}{n}\right)^2\right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_M[f(v_i)] + \frac{2}{n^2} \sum_{i < j} \mathbb{E}_M[f(v_i)f(v_j)] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_M[f(v_i)] - \frac{1}{n^2} \sum_{i < j} (1 - \sigma_{ij}). \end{aligned}$$

Thus the expected variance of naive sampling is

$$\begin{aligned} \mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{naive}})] &= \frac{\mathbb{E}_M[\bar{f}] - \mathbb{E}_M[\bar{f}^2]}{r} \\ &= \frac{\sum_{i \neq j} (1 - \sigma_{ij})}{2n^2 r} = \frac{\sum_{i \neq j} w_{ij}}{2n^2 r} \end{aligned}$$

where  $w_{ij}$  is edge  $(v_i, v_j)$ 's weight in the graph  $G_a$  and  $\sum_{i \neq j} w_{ij}$  is the volume of the graph  $G_a$ .

In the first iteration of greedy partitioning algorithm (Algorithm 2), the randomly generated node sequence is supposed to be  $v_{s_1}, v_{s_2}, \dots, v_{s_n}$ . In the  $k$ -th assignment, the group that  $v_{s_k}$  is assigned to will make the cost function increased the least, thus the increase of cost function should be no more than  $\sum_{l=1}^{k-1} w_{s_k s_l} / r$ .

After the first iteration, the cost function

$$g(\mathcal{P}) \leq \sum_{k=2}^r \sum_{l=1}^{k-1} w_{s_k s_l} / r \leq \frac{1}{r} \sum_{i \neq j} (1 - \sigma_{ij}).$$

Therefore

$$\begin{aligned} &\mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}))] - \mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{naive}}(V, r))] \\ &= \frac{1}{2n^2} g(\mathcal{P}) - \frac{\sum_{i \neq j} (1 - \sigma_{ij})}{2n^2 r} \\ &\leq \frac{\sum_{i \neq j} (1 - \sigma_{ij})}{2n^2 r} - \frac{\sum_{i \neq j} (1 - \sigma_{ij})}{2n^2 r} \\ &= 0. \end{aligned}$$

This finishes the proof.  $\square$

THEOREM 2. *For any non-simple partition  $\mathcal{P}$ , there exists a refined simple partition  $\mathcal{P}'$  of  $\mathcal{P}$ , which can be constructed efficiently, such that partitioned sampling using the refined simple partition  $\mathcal{P}'$  is at least as good as partitioned sampling using the original partition  $\mathcal{P}$ . Specifically,*

$$\mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}'))] \leq \mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}))].$$

PROOF. We first show that the sample variance of partitioned sampling can be written as a weighted summation of the sample variance of naive sampling in each group as follow:

$$\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P})) = \sum_{k=1}^K \frac{n_k^2}{n^2} \text{Var}_S(\hat{f}_{\text{naive}}(V_k, r_k)) \quad (2)$$

where  $n$  is the size of vertex set  $V$ ,  $K$  is the number of groups, and  $n_k$  is the size of  $k$ -th group  $V_k$  of  $\mathcal{P}$ .

According to the definition of partitioned sampling,

$$\hat{f}_{\text{part}}(\mathcal{P}) = \sum_{k=1}^K \frac{n_k}{n} \hat{f}_{\text{naive}}(V_k, r_k).$$

Since the estimate of naive sampling in two different groups are independent, thus for any  $k \neq l$ ,

$$\begin{aligned} \mathbb{E}_S[\hat{f}_{\text{naive}}(V_k, r_k) \hat{f}_{\text{naive}}(V_l, r_l)] \\ = \mathbb{E}_S[\hat{f}_{\text{naive}}(V_k, r_k)] \cdot \mathbb{E}_S[\hat{f}_{\text{naive}}(V_l, r_l)]. \end{aligned}$$

Therefore

$$\begin{aligned} \text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P})) &= \mathbb{E}_S[\hat{f}_{\text{part}}(\mathcal{P})^2] - \mathbb{E}_S[\hat{f}_{\text{part}}(\mathcal{P})]^2 \\ &= \mathbb{E}_S\left[\left(\sum_{k=1}^K \frac{n_k}{n} \hat{f}_{\text{naive}}(V_k, r_k)\right)^2\right] - \mathbb{E}_S\left[\sum_{k=1}^K \frac{n_k}{n} \hat{f}_{\text{naive}}(V_k, r_k)\right]^2 \\ &= \sum_{k=1}^K \frac{n_k^2}{n^2} \mathbb{E}_S[\hat{f}_{\text{naive}}(V_k, r_k)^2] - \sum_{k=1}^K \frac{n_k^2}{n^2} \mathbb{E}_S[\hat{f}_{\text{naive}}(V_k, r_k)]^2 \\ &\quad + \sum_{k \neq l} \frac{n_k n_l}{n^2} \mathbb{E}_S[\hat{f}_{\text{naive}}(V_k, r_k) \hat{f}_{\text{naive}}(V_l, r_l)] \\ &\quad - \sum_{k \neq l} \frac{n_k n_l}{n^2} \mathbb{E}_S[\hat{f}_{\text{naive}}(V_k, r_k)] \mathbb{E}_S[\hat{f}_{\text{naive}}(V_l, r_l)] \\ &= \sum_{k=1}^K \frac{n_k^2}{n^2} \mathbb{E}_S[\hat{f}_{\text{naive}}(V_k, r_k)^2] - \sum_{k=1}^K \frac{n_k^2}{n^2} \mathbb{E}_S[\hat{f}_{\text{naive}}(V_k, r_k)]^2 \\ &= \sum_{k=1}^K \frac{n_k^2}{n^2} \text{Var}_S(\hat{f}_{\text{naive}}(V_k, r_k)). \end{aligned}$$

For any partition  $\mathcal{P}$  with  $K$  groups, if there exists some group  $V_k$  containing more than one sample nodes, according to Lemma 2, we can efficiently find the simple partition  $\mathcal{P}_k^*$  for each group  $V_k$  by greedy partitioning algorithm such that

$$\mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}_k^*))] \leq \mathbb{E}_M[\text{Var}_S(\hat{f}_{\text{naive}}(V_k, r_k))].$$



Thus the refined simple partition  $\mathcal{P}'$  of  $\mathcal{P}$  is constructed by combining all the simple partitions for each group together, and it satisfies that

$$\begin{aligned}\mathbb{E}_M \left[ \text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}')) \right] &= \sum_{k=1}^K \frac{n_k^2}{n^2} \mathbb{E}_M \left[ \text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P}_k^*)) \right] \\ &\leq \sum_{k=1}^K \frac{n_k^2}{n^2} \mathbb{E}_M \left[ \text{Var}_S(\hat{f}_{\text{naive}}(V_k, r_k)) \right] \\ &= \mathbb{E}_M \left[ \text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P})) \right].\end{aligned}$$

This finishes the proof.  $\square$

LEMMA 3. When  $p_i > 0$  for all  $i \in [n]$ , the VIO model has a unique joint distribution for the final expressed opinions in the steady state.

PROOF. The opinion evolution can be viewed as a Markov chain. Each possible assignment of  $v_1, v_2, \dots, v_n$ 's expressed opinions forms one state and the initial state of the Markov chain is  $(f^{(0)}(v_1), f^{(0)}(v_2), \dots, f^{(0)}(v_n))$ . At each Poisson arrival time, the transition from one state to another represents the change of the opinion assignment. Thus the state space consists of all the states reachable from the initial state. The VIO model has a unique steady state distribution for the final expressed opinions if and only if the Markov chain has a unique stationary distribution. In order to prove the existence of the unique stationary distribution of the Markov chain, we only need to prove that the Markov chain is irreducible and aperiodic [19]. Notice that each state in the state space can be reached from the initial state. Meanwhile, each state in the state space can return to the initial state by all the nodes updating their expressed opinions to the innate opinion one by one, which happens with a positive probability. This means that any two states in the state space are connected, indicating the irreducibility of the Markov chain. In addition, the initial state is aperiodic since it has a self-loop in the state transition graph (with probability at least  $\sum_{i=1}^n p_i/n > 0$ ). An irreducible Markov chain is aperiodic if there exists one aperiodic state. Therefore, the Markov chain is irreducible and aperiodic, with the unique stationary distribution being reached after long enough time.  $\square$

LEMMA 4. The expected expressed opinion of each node in the steady state is equal to the expected value of innate opinion, namely, for all  $i \in [n]$ ,

$$\mu_i = \mathbb{E}_M[f^{(\infty)}(v_i)] = \mu^{(0)}.$$

PROOF. We prove this lemma by proving a stronger statement: given any  $t \geq 0$ ,  $\mathbb{E}_M[f^{(t)}(v_i)] = \mu^{(0)}$  for all  $i \in [n]$ . Namely, we want to prove that at any time  $t$ , each node's expected expressed opinion is equal to the expected innate opinion. The proof is by induction on time  $t$ .

In the initial state, each node's expressed opinion (also innate opinion) is generated from an i.i.d. distribution with expected value  $\mu^{(0)}$ , and thus the above statement holds.

Now suppose the statement holds before time  $t$ . It still holds before the next Poisson arrival among all nodes. Suppose the next Poisson arrival comes at time  $t_1$  and its corresponding updating node is  $v_i$ . At this Poisson arrival time  $t_1 > t$ ,  $v_i$  updates its expressed opinion based on its innate opinion and one of its neighbors' expressed opinions. Notice

that the expectations of both its innate opinion  $f^{(0)}(v_i)$  and all its neighbors' expressed opinions  $f^{(t_1)}(v_j)$  are equal to  $\mu^{(0)}$  by the inductive assumption, namely,  $\mathbb{E}_M[f^{(0)}(v_i)] = \mu$  and  $\mathbb{E}_M[f^{(t_1)}(v_j)] = \mu$  for all  $v_j \in N_i$ , where  $N_i$  is the set of  $v_i$ 's neighbors. Thus the expectation of  $v_i$ 's updated expressed opinion  $\mathbb{E}_M[f^{(t_1)}(v_i)]$  is still equal to  $\mu$ . Moreover, other nodes' expected expressed opinions remain equal to  $\mu^{(0)}$  upon time  $t_1$ .

Thus by induction, at any time  $t$ , each node's expected expressed opinion is always equal to  $\mu^{(0)}$ .  $\square$

LEMMA 5. For  $i, j, \ell \in [n]$ ,  $\mathbb{P}[\mathcal{I}_{ij}^\ell]$  is the unique solution of the following linear equation system:

$$\mathbb{P}[\mathcal{I}_{ij}^\ell] = \begin{cases} 0, & i = j \neq \ell, \\ 1, & i = j = \ell, \\ \sum_{a=1}^n \frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i+\lambda_j)d_i} \mathbb{P}[\mathcal{I}_{aj}^\ell] \\ \quad + \sum_{b=1}^n \frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i+\lambda_j)d_j} \mathbb{P}[\mathcal{I}_{ib}^\ell], & i \neq j, \end{cases}$$

where  $d_i = \sum_{j=1}^n A_{ij}$  is node  $v_i$ 's weighted out-degree. In addition, matrix  $Q$  is computed by

$$Q = (I - (I - P)D^{-1}A)^{-1}P,$$

where  $P = \text{diag}(p_1, p_2, \dots, p_n)$  and  $D = \text{diag}(d_1, d_2, \dots, d_n)$  are two diagonal matrices, and matrix  $I - (I - P)D^{-1}A$  is invertible when  $p_i > 0$  for all  $i \in [n]$ .

PROOF. (a) Recall from Definition 3 that  $\mathcal{I}_{ij}^\ell$  denotes the event that two random walks starting from  $v_i$  and  $v_j$  at time  $t = \infty$  eventually meet and the first node they meet at is  $v_\ell \in V$ . This event consists of two steps: 1) The walker at  $v_i$  (or  $v_j$ ) moves to one of its neighbor  $v_a$  (or  $v_b$ ) with probability  $(1-p_i)A_{ia}/d_i$  (or  $(1-p_j)A_{jb}/d_j$ ); 2) two random walks starting from  $v_a$  (or  $v_b$ ) and  $v_j$  (or  $v_i$ ) eventually meet and the first node they meet at is  $v_\ell$ . The probability that the walker at  $v_i$  (resp.  $v_j$ ) make a movement is proportional to  $v_i$ 's (resp.  $v_j$ 's) Poisson rate, that is  $\lambda_i/(\lambda_i + \lambda_j)$  (resp.  $\lambda_j/(\lambda_i + \lambda_j)$ ). Thus when  $i \neq j$ ,  $\mathbb{P}[\mathcal{I}_{ij}^\ell]$  can be calculated by the following recursion:

$$\begin{aligned}\mathbb{P}[\mathcal{I}_{ij}^\ell] &= \sum_{a=1}^n \frac{\lambda_i}{\lambda_i + \lambda_j} \frac{(1-p_i)A_{ia}}{d_i} \mathbb{P}[\mathcal{I}_{aj}^\ell] \\ &\quad + \sum_{b=1}^n \frac{\lambda_j}{\lambda_i + \lambda_j} \frac{(1-p_j)A_{jb}}{d_j} \mathbb{P}[\mathcal{I}_{ib}^\ell].\end{aligned}$$

When  $i = j$ ,  $\mathbb{P}[\mathcal{I}_{ij}^\ell]$  is determined by the following boundary conditions:

$$\mathbb{P}[\mathcal{I}_{ij}^\ell] = \begin{cases} 0, & i = j \neq \ell, \\ 1, & i = j = \ell. \end{cases}$$

By combining the recursive equations and the boundary conditions, we have the following linear equations:

$$\mathbb{P}[\mathcal{I}_{ij}^\ell] = \begin{cases} 0, & i = j \neq \ell, \\ 1, & i = j = \ell, \\ \sum_{a=1}^n \frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i+\lambda_j)d_i} \mathbb{P}[\mathcal{I}_{aj}^\ell] \\ \quad + \sum_{b=1}^n \frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i+\lambda_j)d_j} \mathbb{P}[\mathcal{I}_{ib}^\ell], & i \neq j. \end{cases} \quad (3)$$

(The above proof follows the idea in [21].)

Next we show that the linear equations has a unique solution. For a fixed  $\ell$ , the equations for all terms  $\mathbb{P}[\mathcal{I}_{ij}^\ell]$  such

that  $i \neq j$  form a linear sub-system of  $\binom{n}{2}$  variables and  $\binom{n}{2}$  equations. Therefore, we can solve the whole linear system (3) by solving  $n$  separated linear sub-systems. Each linear sub-system corresponds to a value of  $\ell$ , and it can be solved in  $O\left(\binom{n}{2}^3\right) = O(n^6)^8$ , thus the original linear system (3) can be solved in time  $n \cdot O(n^6) = O(n^7)$ , as mentioned in Section 4.3.

Now, we show that there exists the unique solution for each linear sub-system. Each equation in the linear sub-system can be written as

$$\begin{aligned} \mathbb{P}[\mathcal{I}_{ij}^\ell] &= \sum_{a \neq j} \frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i + \lambda_j)d_i} \mathbb{P}[\mathcal{I}_{aj}^\ell] \\ &\quad + \sum_{b \neq i} \frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i + \lambda_j)d_j} \mathbb{P}[\mathcal{I}_{ib}^\ell] \\ &\quad + \frac{\lambda_i(1-p_i)A_{ij}}{(\lambda_i + \lambda_j)d_i} \cdot 1_{j=\ell} + \frac{\lambda_j(1-p_j)A_{ji}}{(\lambda_i + \lambda_j)d_j} \cdot 1_{i=\ell}. \end{aligned}$$

Let  $k = h(i, j) = (i-1)n + j$ , then we have a bijection  $h$  of subscript between integer  $k$  and ordered pair  $(i, j)$  where  $i < j$ . We can write these equations in matrix form:

$$I\vec{x} = M_1\vec{x} + M_2\vec{x} + \vec{b}$$

where  $\vec{x}$  is the  $\binom{n}{2} \times 1$  vector whose  $k$ -th element is  $\mathbb{P}[\mathcal{I}_{ij}^\ell]$ ;  $M_1$  is the  $\binom{n}{2} \times \binom{n}{2}$  matrix whose  $(k, h(a, j))$  entry is  $\frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i + \lambda_j)d_i}$ ;  $M_2$  is the  $\binom{n}{2} \times \binom{n}{2}$  matrix whose  $(k, h(i, b))$  entry is  $\frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i + \lambda_j)d_j}$ ;  $\vec{b}$  is the  $\binom{n}{2} \times 1$  vector whose  $k$ -th element is  $\frac{\lambda_i(1-p_i)A_{ij}}{(\lambda_i + \lambda_j)d_i} \cdot 1_{j=\ell} + \frac{\lambda_j(1-p_j)A_{ji}}{(\lambda_i + \lambda_j)d_j} \cdot 1_{i=\ell}$ .

If  $I - M_1 - M_2$  is non-singular, then each linear sub-system has a unique solution  $(I - M_1 - M_2)^{-1}\vec{b}$ . In fact, for any row  $s$  of  $I - M_1 - M_2$ , let  $(i, j) = h^{-1}(s)$ , and

$$\begin{aligned} \sum_{t \neq s} |I - M_1 - M_2|_{st} &= \sum_{a \neq j} \frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i + \lambda_j)d_i} + \sum_{b \neq i} \frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i + \lambda_j)d_j} \\ &= \frac{\lambda_i(1-p_i)}{(\lambda_i + \lambda_j)} \sum_{a \neq j} \frac{A_{ia}}{d_i} + \frac{\lambda_j(1-p_j)}{(\lambda_i + \lambda_j)} \sum_{b \neq i} \frac{A_{jb}}{d_j} \\ &\leq \frac{\lambda_i(1-p_i)}{(\lambda_i + \lambda_j)} + \frac{\lambda_j(1-p_j)}{(\lambda_i + \lambda_j)} \\ &< \frac{\lambda_i}{(\lambda_i + \lambda_j)} + \frac{\lambda_j}{(\lambda_i + \lambda_j)} \\ &= |I - M_1 - M_2|_{ss}. \end{aligned}$$

Thus,  $I - M_1 - M_2$  is strictly diagonally dominant, and it is non-singular [1]. Since each linear sub-system has one unique solution, the whole linear system (3) also does.

(b) The probability of a walker from  $v_i$  walking to  $v_j$  in one walk step is  $p_{ij}^{(1)} = (1-p_i)A_{ij}/d_i$ . So we have a matrix form  $P_{VV} = (I - P)D^{-1}A$  whose  $(i, j)$  is  $p_{ij}^{(1)}$ . Therefore, the probability of walking from  $v_i$  to  $v_j$  in exactly  $\ell$  steps is the  $(i, j)$  entry of  $(P_{VV})^\ell$ .

By definition of our model, the probability of  $v_j$  walking to  $v_j'$  (being absorbed) is  $p_j$ . Thus the matrix  $Q$  whose  $(i, j)$  entry is the probability of transition from  $v_i$  to  $v_j'$  can be

<sup>8</sup> $n$ -variable linear system can be solved in time  $O(n^3)$ .

calculated by

$$\begin{aligned} Q &= \sum_{\ell=0}^{\infty} (P_{VV})^\ell P = (I - P_{VV})^{-1}P \\ &= (I - (I - P)D^{-1}A)^{-1}P. \end{aligned}$$

Now we show that  $I - P_{VV}$  is invertible. The  $(i, j)$  entry of  $I - P_{VV}$  is

$$\begin{cases} 1, & \text{if } i = j, \\ -\frac{1-p_i}{d_i}A_{ij}, & \text{if } i \neq j. \end{cases}$$

For any row  $i$  of  $I - P_{VV}$ , the sum of absolute values of its non diagonal elements can be written as  $\sum_{j \neq i} \frac{1-p_i}{d_i}A_{ij} = (1-p_i)(1 - \frac{A_{ii}}{d_i})$ , and it is strictly less than the absolute value of  $i$ -th diagonal elements  $|I - P_{VV}|_{ii} = 1$ . Thus  $I - P_{VV}$  is strictly diagonally dominant, and it is non-singular [1].  $\square$

LEMMA 6. For any  $i, j \in [n]$ , opinion correlation  $\rho_{ij}$  in the steady state is equal to the probability that two coalescing random walks starting from  $v_i$  and  $v_j$  at time  $t = \infty$  end at the same absorbing node in  $V'$ . Moreover, opinion correlation  $\rho_{ij}$  can be computed by

$$\begin{aligned} \rho_{ij} &= \text{Cor}_M(f^{(\infty)}(v_i), f^{(\infty)}(v_j)) \\ &= \sum_{k=1}^n Q_{ik}Q_{jk} + \sum_{\ell=1}^n \mathbb{P}[\mathcal{I}_{ij}^\ell] \left(1 - \sum_{k=1}^n Q_{\ell k}^2\right) \end{aligned}$$

where  $\mathcal{I}_{ij}^\ell$  and  $Q$  are defined in Definition 3, and  $\mathbb{P}[\mathcal{I}_{ij}^\ell]$  and  $Q$  are computed by Lemma 5.

PROOF. (a) In this part, we show that the opinion correlation  $\text{Cor}_M(f^{(\infty)}(v_i), f^{(\infty)}(v_j))$  is equal to the probability that two coalescing random walks starting from  $v_i$  and  $v_j$  at time  $t = \infty$  end at the same absorbing node in  $V'$ . In the proof, we split the randomness  $M$  into two parts: We use  $O$  to denote the randomness of innate opinions which are generated by an i.i.d. distribution, and we use  $E$  to denote the randomness from the opinion evolution.

When  $i = j$ , obviously we have  $\rho_{ij} = 1$ . In this case, the two random walks' paths coincide, thus they are absorbed by the same node in  $V'$  with probability 1.

When  $i \neq j$ , according to the definition of correlation,

$$\begin{aligned} \rho_{ij} &= \text{Cor}_M(f^{(\infty)}(v_i), f^{(\infty)}(v_j)) \\ &= \frac{\mathbb{E}_M[f^{(\infty)}(v_i)f^{(\infty)}(v_j)] - \mathbb{E}_M[f^{(\infty)}(v_i)]\mathbb{E}_M[f^{(\infty)}(v_j)]}{\sqrt{\text{Var}_M[f^{(\infty)}(v_i)]\text{Var}_M[f^{(\infty)}(v_j)]}} \\ &= \frac{\mathbb{E}_M[f^{(\infty)}(v_i)f^{(\infty)}(v_j)] - (\mu^{(0)})^2}{\mu^{(0)} - (\mu^{(0)})^2}. \end{aligned} \quad (4)$$

The third equality holds because for any  $i \in [n]$ ,

$$\begin{aligned} \text{Var}_M[f^{(\infty)}(v_i)] &= \mathbb{E}_M[f^{(\infty)}(v_i)^2] - \mathbb{E}_M[f^{(\infty)}(v_i)]^2 \\ &= \mathbb{E}_M[f^{(\infty)}(v_i)] - \mathbb{E}_M[f^{(\infty)}(v_i)]^2 = \mu^{(0)} - (\mu^{(0)})^2. \end{aligned}$$

Next, we need to calculate  $\mathbb{E}_M[f^{(\infty)}(v_i)f^{(\infty)}(v_j)]$ , which is the probability that two random walkers starting from  $v_i$  and  $v_j$  walk to the nodes in  $V'$  whose original innate opinions are 1. This event consists of two cases: Two random walkers move to the same absorbing node, or two distinct absorbing nodes. Thus we can calculate  $\mathbb{E}_M[f^{(\infty)}(v_i)f^{(\infty)}(v_j)]$  by adding them together.

Let  $\mathcal{M}_{i,j}^{p,q}$  be the event that in the coalescing random walks on  $\overline{G}$ , a random walker starting from  $v_i$  is absorbed by  $v'_p$ , while another random walker starting from  $v_j$  is absorbed by  $v'_q$ . Note that  $\mathcal{M}_{i,j}^{p,q}$  is measurable under randomness  $E$ . It only depends on the structure of  $\overline{G}$  and is independent of the initial value in  $V'$ :

$$\mathbb{P}_E \left[ \mathcal{M}_{i,j}^{p,q} \mid f^{(0)}(v_1), f^{(0)}(v_2), \dots, f^{(0)}(v_n) \right] = \mathbb{P}_E \left[ \mathcal{M}_{i,j}^{p,q} \right].$$

Thus  $\mathbb{E}_M[f^{(\infty)}(v_i)f^{(\infty)}(v_j)]$  can be written as:

$$\begin{aligned} & \mathbb{E}_M[f^{(\infty)}(v_i)f^{(\infty)}(v_j)] \\ &= \mathbb{E}_{O,E}[f^{(\infty)}(v_i)f^{(\infty)}(v_j)] \\ &= \sum_{p \neq q} \left( \mathbb{P}_{O,E} \left[ \mathcal{M}_{i,j}^{p,q} \mid f^{(0)}(v_p)f^{(0)}(v_q) = 1 \right] \right. \\ & \quad \cdot \mathbb{P}_{O,E} \left[ f^{(0)}(v_p)f^{(0)}(v_q) = 1 \right] \\ & \quad + \sum_{p=1}^n \mathbb{P}_{O,E} \left[ \mathcal{M}_{i,j}^{p,p} \mid f^{(0)}(v_p) = 1 \right] \mathbb{P}_{O,E} \left[ f^{(0)}(v_p) = 1 \right] \\ &= \sum_{p \neq q} \mathbb{P}_E \left[ \mathcal{M}_{i,j}^{p,q} \right] \mathbb{P}_O \left[ f^{(0)}(v_p)f^{(0)}(v_q) = 1 \right] \\ & \quad + \sum_{p=1}^n \mathbb{P}_E \left[ \mathcal{M}_{i,j}^{p,p} \right] \mathbb{P}_O \left[ f^{(0)}(v_p) = 1 \right] \\ &= (\mu^{(0)})^2 \sum_{p \neq q} \mathbb{P}_E \left[ \mathcal{M}_{i,j}^{p,q} \right] + \mu^{(0)} \sum_{p=1}^n \mathbb{P}_E \left[ \mathcal{M}_{i,j}^{p,p} \right] \\ &= \mu^{(0)} \sum_{p=1}^n \mathbb{P}_E \left[ \mathcal{M}_{i,j}^{p,p} \right] + (\mu^{(0)})^2 \left( 1 - \sum_{p=1}^n \mathbb{P}_E \left[ \mathcal{M}_{i,j}^{p,p} \right] \right) \\ &=: \mu^{(0)} p_{\text{same}}(i, j) + (\mu^{(0)})^2 (1 - p_{\text{same}}(i, j)). \end{aligned}$$

In the last equation, we use  $p_{\text{same}}(i, j)$  to denote the probability that two coalescing random walks starting from  $v_i$  and  $v_j$  end at the same node in  $V'$ . Thus

$$\begin{aligned} \rho_{ij} &= \frac{\mathbb{E}_M[f^{(\infty)}(v_i)f^{(\infty)}(v_j)] - (\mu^{(0)})^2}{\mu^{(0)} - (\mu^{(0)})^2} \\ &= \frac{\left[ \mu^{(0)} p_{\text{same}}(i, j) + (\mu^{(0)})^2 (1 - p_{\text{same}}(i, j)) \right] - (\mu^{(0)})^2}{\mu^{(0)} - (\mu^{(0)})^2} \\ &= p_{\text{same}}(i, j). \end{aligned}$$

This means that the opinion correlation  $\rho_{ij}$  is equal to the probability that two coalescing random walks starting from  $v_i$  and  $v_j$  end at the same absorbing node in  $V'$ .

(b) We now calculate  $p_{\text{same}}$  in this part. Let  $\mathcal{H}_{ij}^k$  be the event that two coalescing random walks starting from  $v_i$  and  $v_j$  are both absorbed by node  $v'_k$  without meeting each other at a node in  $V$ . According to the definitions of events  $\mathcal{M}_{i,j}^{p,q}$ ,  $\mathcal{I}_{ij}^\ell$  and  $\mathcal{H}_{ij}^k$ , we have

$$\mathbb{P}_E \left[ \mathcal{M}_{i,j}^{k,k} \right] = \sum_{\ell=1}^n \mathbb{P} \left[ \mathcal{I}_{ij}^\ell \right] Q_{\ell k} + \mathbb{P} \left[ \mathcal{H}_{ij}^k \right] \quad (5)$$

where  $Q_{\ell k}$  is the probability that a random walker starting from node  $v_\ell$  at time ends at  $v'_k \in V'$ .

Notice that  $Q_{ik}Q_{jk}$  represents the probability that two *non*-coalescing random walks starting from  $v_i$  and  $v_j$  end at

node  $v'_k$ , thus it can be written as

$$Q_{ik}Q_{jk} = \sum_{\ell=1}^n \mathbb{P} \left[ \mathcal{I}_{ij}^\ell \right] Q_{\ell k}^2 + \mathbb{P} \left[ \mathcal{H}_{ij}^k \right]. \quad (6)$$

Combining Equation (5) and (6),

$$\mathbb{P}_E \left[ \mathcal{M}_{i,j}^{k,k} \right] = \sum_{\ell=1}^n \mathbb{P} \left[ \mathcal{I}_{ij}^\ell \right] (Q_{\ell k} - Q_{\ell k}^2) + Q_{ik}Q_{jk}. \quad (7)$$

Therefore,

$$\begin{aligned} \rho_{ij} &= p_{\text{same}}(i, j) = \sum_{k=1}^n \mathbb{P}_E \left[ \mathcal{M}_{i,j}^{k,k} \right] \\ &= \sum_{k=1}^n \left( \sum_{\ell=1}^n \mathbb{P} \left[ \mathcal{I}_{ij}^\ell \right] (Q_{\ell k} - Q_{\ell k}^2) + Q_{ik}Q_{jk} \right) \\ &= \sum_{\ell=1}^n \mathbb{P} \left[ \mathcal{I}_{ij}^\ell \right] \left( 1 - \sum_{k=1}^n Q_{\ell k}^2 \right) + \sum_{k=1}^n Q_{ik}Q_{jk}. \end{aligned}$$

This finishes the proof.  $\square$

**THEOREM 3.** For any two nodes  $v_i$  and  $v_j$ , their opinion similarity  $\sigma_{ij}$  in the steady state of the VIO model is equal to:

$$\sigma_{ij} = 1 - 2\mu^{(0)}(1 - \mu^{(0)})(1 - \rho_{ij})$$

where opinion correlation  $\rho_{ij}$  is computed by Lemma 6.

**PROOF.** Recall Eq. (1),

$$\begin{aligned} & \mathbb{E}_M[f^{(\infty)}(v_i)f^{(\infty)}(v_j)] \\ &= \frac{1}{2}(\sigma_{ij} + \mathbb{E}_M[f^{(\infty)}(v_i)] + \mathbb{E}_M[f^{(\infty)}(v_j)] - 1) \\ &= \frac{1}{2}(\sigma_{ij} - 1) + \mu^{(0)}, \end{aligned}$$

and Eq. (4),

$$\mathbb{E}_M[f^{(\infty)}(v_i)f^{(\infty)}(v_j)] = \mu^{(0)}(1 - \mu^{(0)})\rho_{ij} + (\mu^{(0)})^2.$$

Thus we can obtain

$$\sigma_{ij} = 1 - 2\mu^{(0)}(1 - \mu^{(0)})(1 - \rho_{ij}).$$

This finishes the proof.  $\square$

## B. OBJECTIVE FUNCTION OF THE OPS PROBLEM

In the definition of the OPS problem (Definition 1), we use the expected variance  $\mathbb{E}_M[\text{Var}_S(\hat{f})]$  as the objective function. Another intuitive setting of the objective function is  $\text{Var}_{M,S}(\hat{f})$  which combine all the randomness into the variance together. We now show that these two objective functions are equivalent. Actually,

$$\begin{aligned} \text{Var}_{M,S}(\hat{f}) &= \mathbb{E}_{M,S}[\hat{f}^2] - \mathbb{E}_{M,S}[\hat{f}]^2 \\ &= \mathbb{E}_M \mathbb{E}_S[\hat{f}^2] - (\mathbb{E}_M \mathbb{E}_S[\hat{f}])^2. \end{aligned}$$

Due to the unbiasedness of the estimator  $\hat{f}$ , we have

$$\mathbb{E}_S[\hat{f}] = \bar{f}.$$

Thus

$$\text{Var}_S(\hat{f}) = \mathbb{E}_S[\hat{f}^2] - \mathbb{E}_S[\hat{f}]^2 = \mathbb{E}_S[\hat{f}^2] - \bar{f}^2.$$

Therefore

$$\begin{aligned}\text{Var}_{M,S}(\hat{f}) &= \mathbb{E}_M \mathbb{E}_S[\hat{f}^2] - (\mathbb{E}_M \mathbb{E}_S[\hat{f}])^2 \\ &= \mathbb{E}_M [\text{Var}_S(\hat{f}) + \bar{f}^2] - (\mathbb{E}_M[\bar{f}])^2 \\ &= \mathbb{E}_M[\text{Var}_S(\hat{f})] + \text{Var}_M(\bar{f}).\end{aligned}$$

Since  $\text{Var}_M(\bar{f})$  stays constant with different partition, objective functions  $\text{Var}_{M,S}(\hat{f})$  and  $\mathbb{E}_M[\text{Var}_S(\hat{f})]$  are equivalent for the OPS problem.

In the definition of OPS problem (Definition 1), we use  $\mathbb{E}_M[\text{Var}_S(\hat{f})]$  as our objective function where  $M$  represents the randomness from the evolution model and  $S$  represents the randomness from sampling. Since our partitioning algorithms (i.e., Algorithm 1 and 2) are randomized algorithms, the randomness  $S$  can further be divided into two parts: the randomness from partitioning algorithms (denoted as  $P$ ) and the randomness from sample selecting in each component (denoted as  $C$ ). Thus the objective function of our OPS problem can be specifically written as  $\mathbb{E}_M[\text{Var}_{P,C}(\hat{f})]$ . However, in our experimental evaluation (Section 5), we compute  $\mathbb{E}_M[\text{Var}_C(\hat{f})]$  for each partition and take average of them to be the “ $\mathbb{E}_M[\text{Var}_S(\hat{f})]$ ”, which is strictly as  $\mathbb{E}_P \mathbb{E}_M[\text{Var}_C(\hat{f})]$ . In fact, they are the equivalent as we show below. Notice that the randomness  $M$  and  $P$  are independent, thus  $\mathbb{E}_M$  and  $\mathbb{E}_P$  are commutative. Thus

$$\begin{aligned}\mathbb{E}_P \mathbb{E}_M[\text{Var}_C(\hat{f})] &= \mathbb{E}_M[\text{Var}_S(\hat{f})] \\ &= \mathbb{E}_P \mathbb{E}_M [\mathbb{E}_C[\hat{f}^2] - \mathbb{E}_C[\hat{f}]^2] \\ &\quad - \mathbb{E}_M [\mathbb{E}_P \mathbb{E}_C[\hat{f}^2] - \mathbb{E}_P \mathbb{E}_C[\hat{f}]^2] \\ &= \mathbb{E}_M \mathbb{E}_P \mathbb{E}_C[\hat{f}^2] - \mathbb{E}_M \mathbb{E}_P [\mathbb{E}_C[\hat{f}]^2] \\ &\quad - \mathbb{E}_M \mathbb{E}_P \mathbb{E}_C[\hat{f}^2] + \mathbb{E}_M [\mathbb{E}_P \mathbb{E}_C[\hat{f}]^2] \\ &= \mathbb{E}_M \left[ \left( \mathbb{E}_P \mathbb{E}_C[\hat{f}] \right)^2 - \mathbb{E}_P [\mathbb{E}_C[\hat{f}]^2] \right] \\ &= \mathbb{E}_M \text{Var}_P (\mathbb{E}_C[\hat{f}]) \\ &= \mathbb{E}_M \text{Var}_P(\bar{f}) \\ &= 0.\end{aligned}\tag{8}$$

Equation (8) holds because  $\hat{f}$  is an unbiased estimate for any partition  $\mathcal{P}$ . Thus  $\mathbb{E}_P \mathbb{E}_M[\text{Var}_C(\hat{f})]$  and  $\mathbb{E}_M[\text{Var}_S(\hat{f})]$  are the same, and we do not distinguish them in the paper.

### C. SDP PARTITIONING ALGORITHM

We now present the formulation of our SDP partitioning algorithm for solving the Max- $r$ -Cut problem for the graph  $G_a$ . The task is to find  $r$  groups  $V_1, V_2, \dots, V_r$  in order to maximize the following function:

$$\sum_{k \neq l} \text{Cut}_{G_a}(V_k, V_l)$$

where  $\text{Cut}_{G_a}(V_k, V_l)$  is defined by  $\sum_{v_i \in V_k, v_j \in V_l} w_{ij}$ .

Frieze and Jerrum [8] propose an approximation algorithm for Max- $r$ -Cut using Semi-Definite Programming (SDP) as a relaxation. We adopt their algorithm for solving our problem as follows.

Take an equilateral simplex in  $\mathbb{R}^{r-1}$  with vertices  $\vec{b}_1, \vec{b}_2, \dots, \vec{b}_r$ . Let  $\vec{c} = (\vec{b}_1 + \vec{b}_2 + \dots + \vec{b}_r)/r$ , and let  $\vec{a}_k = \frac{\vec{b}_k - \vec{c}}{\|\vec{b}_k - \vec{c}\|}$  for  $1 \leq k \leq r$ . Thus  $\vec{a}_1, \vec{a}_2, \dots, \vec{a}_r$  have the following property:

$$\vec{a}_k \cdot \vec{a}_l = \begin{cases} 1, & \text{if } \vec{a}_k = \vec{a}_l; \\ -\frac{1}{r-1}, & \text{if } \vec{a}_k \neq \vec{a}_l. \end{cases}$$

We use  $\vec{y}_i \in \{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_r\}$  to present which group node  $v_i$  is located in. If node  $v_i$  is in  $k$ -th group, then  $\vec{y}_i = \vec{a}_k$ . In this way, the maximization problem can be written as

$$\begin{aligned}\text{Maximize} \quad & \frac{r-1}{r} \sum_{i \neq j} w_{ij} (1 - \vec{y}_i \cdot \vec{y}_j) \quad (\text{IP}) \\ \text{Subject to} \quad & \vec{y}_i \in \{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_r\}, i \in \{1, 2, \dots, n\}.\end{aligned}$$

To obtain the SDP relaxation, we replace  $\vec{y}_i \cdot \vec{y}_j$  by  $(i, j)$ -entry of the positive semi-definite symmetric matrix  $Y$  whose diagonal elements are equal to 1, and relax  $\vec{y}_i \cdot \vec{y}_j$  to be not less than  $-\frac{1}{r-1}$ .

$$\begin{aligned}\text{Maximize} \quad & \frac{r-1}{r} \sum_{i \neq j} w_{ij} (1 - Y_{ij}) \quad (\text{SDP}) \\ \text{Subject to} \quad & Y_{ii} = 1, \forall i, \\ & Y \succeq 0, \\ & Y_{ij} \geq -\frac{1}{r-1}, \forall i \neq j, \\ & Y \text{ is symmetric.}\end{aligned}$$

Our SDP partitioning algorithm is performed by solving the above SDP problem and rounding the SDP-relaxed solution to IP-flexible solution, which is shown in Algorithm 1.

### D. EFFICIENT CORRELATION COMPUTATION

Naive correlation computation directly using Lemma 5 and 6 by solving the linear equation system for  $\{\mathbb{P}[\mathcal{I}_{ij}^k]\}$ 's would have a running time of  $O(n^7)$  (See in proof of Lemma 5). We now improve the running time to  $O(nmR)$  by a carefully designed iterative computation method, where  $m$  is the number of edges of the social graph  $G$  and  $R$  is the number of iterations.

We first consider the event that a walker starting at  $v_i$  is absorbed at  $v'_j \in V'$  after  $k$  random walk steps in the coalescing random walk model. We use  $q_{ij}^{(k)}$  to represent the probability of the above event happening. Initially, the walker is located at  $v_i$ . With probability  $p_i$ , she takes her first step to the sink node  $v'_i$ . Then the event that she is absorbed by  $v'_j \in V'$  after  $k$  steps happens iff  $i = j$ . If she does not walk to the sink node in her first step, but walk to one of her neighbors  $v_a$  (happening with probability  $(1 - p_i)A_{ia}/d_i$ ), then the event happens iff she walks from  $v_a$  to  $v'_j$  in her rest  $k-1$  steps. Thus  $q_{ij}^{(k)}$  can be computed iteratively by:

$$q_{ij}^{(k)} = p_i \cdot 1_{i=j} + \sum_{a=1}^n \frac{(1 - p_i)A_{ia}}{d_i} q_{aj}^{(k-1)}.$$

The running time of computing  $\{q_{ij}^{(k)}\}$ 's with one iteration is

$$\sum_{i=1}^n \sum_{j=1}^n (1 + d_i) = O(nm),$$



where  $m$  is number of edges of the social graph  $G$ . We remark that when  $k \rightarrow \infty$ ,  $q_{ij}^{(k)}$  approaches to the  $(i, j)$ -entry of matrix  $Q$  defined in Definition 3. Thus  $Q$  can be computed in the above iterative way. We further remark that  $Q$  can be computed column by column (fixing subscript  $j$ ) to save running space.

Now we consider two walkers take coalescing random walks on the graph  $\bar{G}$  starting at  $v_i$  and  $v_j$  respectively. We use notation  $M_{ij}^{(k)}$  to denote the probability that their walks coalesce after they taking altogether  $k$  random walk steps. If  $i = j$ , two walkers have coalesced since the beginning, thus  $M_{ij}^{(k)}$  is always equal to one. When  $i \neq j$ , with probability  $\frac{\lambda_i}{\lambda_i + \lambda_j}$ , the first walk step is taken by the walker starting at  $v_i$ . If she walks to her sink node  $v'_i$  (happening with probability  $p_i$ ), then the other walker who is at  $v_j$  must walk to the same sink node  $v'_i$  in  $k-1$  steps (happening with probability  $q_{ji}^{(k-1)}$ ). If she does not walk to her sink node but one of her neighbors  $v_a$  (happening with probability  $(1-p_i)A_{ia}/d_i$ ), then two walkers will coalesce in the rest  $k-1$  steps with probability  $M_{aj}^{(k-1)}$ . The case that the first step is taken by the walker starting at  $v_j$  follows the similar analysis. Thus  $M_{ij}^{(k)}$  can be calculated by:

$$M_{ij}^{(k)} = \frac{\lambda_i}{\lambda_i + \lambda_j} \left[ p_i q_{ji}^{(k-1)} + \sum_{a=1}^n \frac{(1-p_i)A_{ia}}{d_i} M_{aj}^{(k-1)} \right] + \frac{\lambda_j}{\lambda_i + \lambda_j} \left[ p_j q_{ij}^{(k-1)} + \sum_{a=1}^n \frac{(1-p_j)A_{ja}}{d_j} M_{ai}^{(k-1)} \right].$$

The running time of computing  $\{M_{ij}^{(k)}\}$ 's with one iteration is

$$\sum_{i=1}^n \sum_{j=1}^n (1 + d_i) + (1 + d_j) = O(nm),$$

where  $m$  is number of edges of the social graph  $G$ .

According to Lemma 6, opinion correlation  $\rho_{ij}$  in the steady state is equal to  $\lim_{k \rightarrow \infty} M_{ij}^{(k)}$ . Thus we can obtain people's opinion correlations by computing  $\{M_{ij}^{(k)}\}$ 's and  $\{q_{ij}^{(k)}\}$ 's iteratively in time  $O(nmR)$  where  $R$  is the number of iterations. We remark that for any  $i, j \in [n]$ ,  $M_{ij}^{(k)}$  and  $q_{ij}^{(k)}$  monotonically increase with increasing  $k$ , and both have the upper bound 1. Thus the above iterations will converge.

## E. BALANCED PARTITION

A partition is called the *balanced* partition if all the groups have equal sizes. We now show that partitioned sampling using balanced simple partitions is always at least as good as naive sampling.

LEMMA 7. *Given a vertex set  $V$  with  $n$  nodes and sample size  $r < n$ , partitioned sampling using any balanced simple partition  $\mathcal{P}$  of  $V$ , is at least as good as naive sampling. Specifically,*

$$\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P})) \leq \text{Var}_S(\hat{f}_{\text{naive}}(V, r)).$$

PROOF. It has been known that, the sample variance of naive sampling is

$$\text{Var}_S(\hat{f}_{\text{naive}}(V, r)) = \frac{1}{r} \bar{f}(1 - \bar{f})$$

where  $\bar{f}$  is the average opinion of the entire population.

Recall Eq. (2), we have

$$\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P})) = \sum_{k=1}^r \frac{n_k^2}{n^2} \text{Var}_S(\hat{f}_{\text{naive}}(V_k, 1)).$$

For the balanced partition  $\mathcal{P}$ ,

$$\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P})) = \frac{1}{r^2} \sum_{k=1}^r \bar{f}_k(1 - \bar{f}_k)$$

where  $\bar{f}_k$  is the average opinion of the  $k$ -th group.

Notice that  $\bar{f} = \sum_{k=1}^r \bar{f}_k / r$ , thus we have

$$\begin{aligned} & \text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P})) - \text{Var}_S(\hat{f}_{\text{naive}}(V, r)) \\ &= \frac{1}{r^2} \sum_{k=1}^r \bar{f}_k(1 - \bar{f}_k) - \frac{1}{r} \bar{f}(1 - \bar{f}) \\ &= \frac{1}{r^2} \sum_{k=1}^r \bar{f}_k(1 - \bar{f}_k) - \frac{1}{r^2} \left( \sum_{k=1}^r \bar{f}_k \right) \left( 1 - \frac{1}{r} \sum_{k=1}^r \bar{f}_k \right) \\ &= -\frac{1}{r^3} \left[ r \sum_{k=1}^r \bar{f}_k^2 - \left( \sum_{k=1}^r \bar{f}_k \right)^2 \right]. \end{aligned}$$

According to Cauchy-Schwarz inequality,

$$\left( \sum_{k=1}^r 1^2 \right) \cdot \left( \sum_{k=1}^r \bar{f}_k^2 \right) \geq \left( \sum_{k=1}^r 1 \cdot \bar{f}_k \right)^2.$$

Thus we have

$$\text{Var}_S(\hat{f}_{\text{part}}(\mathcal{P})) - \text{Var}_S(\hat{f}_{\text{naive}}(V, r)) \leq 0.$$

This finishes the proof.  $\square$

According to Lemma 7, if we make a small modification of the partitioning algorithms to force the output partition to be a balanced partition, the balanced partitioning algorithms always achieve better sampling quality than naive sampling, no matter whether the opinion similarity estimation is accurate or not. For example, for the greedy partitioning (Algorithm 2), we set each group's capacity to be  $n/r$ . When doing the greedy assignment, if some group is full, then we force to assign the rest ungrouped nodes to the groups which are not full.

## F. VIO MODEL WITH NON-IDENTICAL INNATE OPINION DISTRIBUTION

In the VIO model (section 4), we regard the individuals' innate opinions as generated from an i.i.d. distribution. However, this assumption may violate in some cases, say different people may have different expected innate opinions. For these cases, we need to relax this assumption of the VIO model. We still assume that the individuals' innate opinions are independent, but each person  $v_i$  has her own expected innate opinion  $\mathbb{E}[f^{(0)}(v_i)] = \mu_i^{(0)}$ . In order to distinguish the source of the randomness, we split the randomness  $M$  into two parts: We use  $O$  to denote the randomness of innate opinions which are generated by the Bernoulli distribution, and use  $E$  to denote the randomness from the opinion evolution (These notations are first introduced in the proof of Lemma 6). Thus  $\mathbb{E}[f^{(0)}(v_i)] = \mu_i^{(0)}$  is clarified as  $\mathbb{E}_O[f^{(0)}(v_i)] = \mu_i^{(0)}$ .

In this section, we will give a sketchy analysis of the VIO model with non-identical innate opinion distribution. We still use notation  $\mu_i$  to denote  $v_i$ 's expected expressed opinion when the evolution converges, that is  $\mu_i = \mathbb{E}_M[f^{(\infty)}(v_i)]$ . Notice that the computation method of  $Q$  and  $\{\mathbb{P}[\mathcal{I}_{ij}^\ell]\}$ 's (Lemma 5) still holds for the VIO model with non-identical innate opinion distribution.

According to the definition of matrix  $Q$  (Definition 3), it is easy to verify that for any  $i \in [n]$ ,

$$\mathbb{E}_E[f^{(\infty)}(v_i)] = \sum_{j=1}^n Q_{ij} f^{(0)}(v_j).$$

Thus

$$\mu_i = \mathbb{E}_O \mathbb{E}_E[f^{(\infty)}(v_i)] = \sum_{j=1}^n Q_{ij} \mathbb{E}_O[f^{(0)}(v_j)] = \sum_{j=1}^n Q_{ij} \mu_j^{(0)}.$$

Next, we focus on computing the opinion similarity between any pair of nodes  $(v_i, v_j)$ , which is the probability that two random walkers starting from  $v_i$  and  $v_j$  have the same final expressed opinions. Recall Eq. (1) that

$$\sigma_{ij} = 2 \mathbb{E}_M[f^{(\infty)}(v_i) f^{(\infty)}(v_j)] + 1 - \mu_i - \mu_j.$$

In order to obtain  $\sigma_{ij}$ , we compute  $\mathbb{E}_M[f^{(\infty)}(v_i) f^{(\infty)}(v_j)]$ , which is the probability that two random walkers starting from  $v_i$  and  $v_j$  finally walk to the nodes in  $V'$  whose innate opinions are equal to one.

Following the notations used in the proof of Lemma 6, let  $\mathcal{M}_{i,j}^{p,q}$  be the event that in the coalescing random walks on  $\overline{G}$ , a random walker starting from  $v_i$  is absorbed by  $v'_p$ , while another random walker starting from  $v_j$  is absorbed by  $v'_q$ . Note that  $\mathcal{M}_{i,j}^{p,q}$  is measurable under randomness  $E$ . Thus

$$\begin{aligned} \mathbb{E}_M[f^{(\infty)}(v_i) f^{(\infty)}(v_j)] &= \mathbb{P}_M[f^{(\infty)}(v_i) = f^{(\infty)}(v_j) = 1] \\ &= \sum_{p=1}^n \mathbb{P}_E[\mathcal{M}_{i,j}^{p,p}] \mathbb{P}_O[f^{(0)}(v_p) = 1] \\ &\quad + \sum_{p \neq q} \mathbb{P}_E[\mathcal{M}_{i,j}^{p,q}] \mathbb{P}_O[f^{(0)}(v_p) = 1, f^{(0)}(v_q) = 1] \\ &= \sum_{p=1}^n \mathbb{P}_E[\mathcal{M}_{i,j}^{p,p}] \mu_p^{(0)} + \sum_{p \neq q} \mathbb{P}_E[\mathcal{M}_{i,j}^{p,q}] \mu_p^{(0)} \mu_q^{(0)}. \end{aligned}$$

Recall Eq. (7)

$$\mathbb{P}_E[\mathcal{M}_{i,j}^{p,p}] = \sum_{\ell=1}^n \mathbb{P}[\mathcal{I}_{ij}^\ell] (Q_{\ell p} - Q_{\ell p}^2) + Q_{ip} Q_{jp},$$

and use the same technique of Eq. (6)

$$\mathbb{P}_E[\mathcal{M}_{i,j}^{p,q}] = Q_{ip} Q_{jq} - \sum_{\ell=1}^n \mathbb{P}[\mathcal{I}_{ij}^\ell] Q_{\ell p} Q_{\ell q}.$$

Therefore, after some calculations,

$$\begin{aligned} \mathbb{E}_M[f^{(\infty)}(v_i) f^{(\infty)}(v_j)] &= \sum_{p=1}^n \mathbb{P}_E[\mathcal{M}_{i,j}^{p,p}] \mu_p^{(0)} + \sum_{p \neq q} \mathbb{P}_E[\mathcal{M}_{i,j}^{p,q}] \mu_p^{(0)} \mu_q^{(0)} \\ &= \mu_i \mu_j + \sum_{p=1}^n \mu_p^{(0)} (1 - \mu_p^{(0)}) Q_{ip} Q_{jp} \\ &\quad + \sum_{\ell=1}^n \mathbb{P}[\mathcal{I}_{ij}^\ell] \left( \mu_\ell (1 - \mu_\ell) - \sum_{p=1}^n \mu_p^{(0)} (1 - \mu_p^{(0)}) Q_{\ell p}^2 \right). \end{aligned}$$

Thus

$$\begin{aligned} \sigma_{ij} &= (1 - \mu_i)(1 - \mu_j) + \mu_i \mu_j + 2 \sum_{p=1}^n \mu_p^{(0)} (1 - \mu_p^{(0)}) Q_{ip} Q_{jp} \\ &\quad + 2 \sum_{\ell=1}^n \mathbb{P}[\mathcal{I}_{ij}^\ell] \left( \mu_\ell (1 - \mu_\ell) - \sum_{p=1}^n \mu_p^{(0)} (1 - \mu_p^{(0)}) Q_{\ell p}^2 \right). \end{aligned}$$

With opinion similarities calculated, then we can use partitioned sampling method described in Section 3 to do efficient sampling.

## G. SIGNED VIO MODEL

In this section, we provide a sketchy analysis of the signed VIO model, which allows negative edges in the graph. Given a weighted directed social graph  $G = (V, A)$  where  $A$  is the weighted adjacency matrix with  $A_{ij} \neq 0$  if and only if edge  $(v_i, v_j)$  exists, with  $A_{ij} \in \mathbb{R}^+ \cup \mathbb{R}^-$  as the weight of edge  $(v_i, v_j)$ . At time 0, each node  $v_i$  generates its innate opinion  $f^{(0)}(v_i)$  from an i.i.d. distribution with mean  $\mu^{(0)}$ . When  $v_i$ 's Poisson arrives, it sets its expressed opinion  $f^{(t)}(v_i)$  to be its own innate opinion  $f^{(0)}(v_i)$  with an inward probability  $p_i$ , or with probability  $1 - p_i$  node  $v_i$  randomly selects a neighbor  $v_j$  of  $v_i$  with probability proportional to absolute value of the weight of the edge  $(v_i, v_j)$ , i.e., with probability  $(1 - p_i)|A_{ij}| / \sum_{k=1}^n |A_{ik}|$ , and sets  $f^{(t)}(v_i)$  to  $v_j$ 's expressed opinion  $f^{(t)}(v_j)$  when  $A_{ij} > 0$  or the opposite of  $v_j$ 's expressed opinion  $1 - f^{(t)}(v_j)$  when  $A_{ij} < 0$ . Similar to the VIO model, when  $p_i > 0$  for all  $i \in [n]$ , the signed VIO model has a unique steady state distribution for the final expressed opinions. For simplicity, we use  $f_i$  (and  $f_i(t)$ ) to denote  $f^{(\infty)}(v_i)$  (and  $f^{(t)}(v_i)$ ). We define  $f'_i$  (and  $f'_i(t)$ ) to be  $2f_i - 1$  (and  $2f_i(t) - 1$ ). Thus  $f'_i(t) = 1$  if  $f_i(t) = 1$ , and  $f'_i(t) = -1$  if  $f_i(t) = 0$ . We list all the notations used in this section in Table 1. In order to distinguish the source of the randomness, we split the randomness  $M$  into two parts: We use  $O$  to denote the randomness of innate opinions which are generated by the Bernoulli distribution, and use  $E$  to denote the randomness from the opinion evolution (These notations are first introduced in the proof of Lemma 6).

The expected final expressed opinions can be calculated by (similar to the computation of  $Q$  in the VIO model)

$$\begin{aligned} \mathbb{E}_E[\vec{f}] &= \sum_{k=0}^{\infty} ((I - P)D^{-1}A)^k P \vec{f}(0) \\ &= (I - (I - P)D^{-1}A) P \vec{f}(0) \end{aligned}$$

where  $\vec{f} = (f'_1, \dots, f'_n)^T$  and  $\vec{f}(0) = (f'_1(0), \dots, f'_n(0))^T$ . Notice that the  $\mathbb{E}_E[\vec{f}]$  can also be written as  $(Q^+ - Q^-) \vec{f}(0)$

| Notation   | Representation   |
|--|--|
| $A^+, A^-, \bar{A}$                                | $A^+$ (resp. $A^-$ ) is the non-negative adjacency matrix representing positive (resp. negative) edges of $G$ , with $A = A^+ - A^-$ and $\bar{A} = A^+ + A^-$ .   |
| $Q^+, Q^-$   | The $(i, j)$ -entry of $Q^+$ (resp. $Q^-$ ) is the probability that the random walk starting from $v_i$ is absorbed by $v'_j$ in $V'$ while the number of walking steps is even (resp. odd).   |
| $p_{\text{same}}^+(i, j), p_{\text{same}}^-(i, j)$ | $p_{\text{same}}^+(i, j)$ (resp. $p_{\text{same}}^-(i, j)$ ) is the probability that two random coalescing walks starting from $v_i$ and $v_j$ are absorbed by the same node in $V'$ while the summation of two walks' steps is even (resp. odd).  |
| $p_{\text{diff}}^+(i, j), p_{\text{diff}}^-(i, j)$ | $p_{\text{diff}}^+(i, j)$ (resp. $p_{\text{diff}}^-(i, j)$ ) is the probability that two random coalescing walks starting from $v_i$ and $v_j$ are absorbed by different nodes in $V'$ while the summation of two walks' steps is even (resp. odd).  |
| $f_i(t), f'_i(t), f_i, f'_i$                       | $f_i(t)$ and $f'_i(t)$ are both representing $v_i$ 's expressed opinion at time $t$ . The difference is $f_i(t) \in \{0, 1\}$ and $f'_i(t) \in \{-1, 1\}$ . $f'_i(t)$ can be obtained by $2f_i(t) - 1$ . $f_i$ and $f'_i$ represent $v_i$ 's final expressed opinion ( $t \rightarrow \infty$ ). |
| $\mathcal{I}_{ij}^{l+}, \mathcal{I}_{ij}^{l-}$     | $\mathcal{I}_{ij}^{l+}$ (resp. $\mathcal{I}_{ij}^{l-}$ ) is the event that two random walks starting from $v_i$ and $v_j$ eventually meet and the first node they meet at is $v_l \in V$ while the summation of the steps they have taken is even (resp. odd).                                   |
| $\mathcal{H}_{ij}^{k+}, \mathcal{H}_{ij}^{k-}$     | $\mathcal{H}_{ij}^{k+}$ (resp. $\mathcal{H}_{ij}^{k-}$ ) is the event that two coalescing random walks starting from $v_i$ and $v_j$ are both absorbed by node $v'_k$ without meeting each other at a node in $V$ while the summation of two walks' steps is even (resp. odd).                   |

**Table 1: Notations**

for any  $\vec{f}'(0)$ , so we have

$$Q^+ - Q^- = (I - (I - P)D^{-1}A)^{-1}P.$$

In the analysis of the unsigned VIO model, we have

$$Q^+ + Q^- = (I - (I - P)D^{-1}\bar{A})^{-1}P.$$

Therefore, we can obtain

$$Q^+ = (I - (I - P)D^{-1}A^+)^{-1}P,$$

and

$$Q^- = (I - (I - P)D^{-1}A^-)^{-1}P.$$

Notice that for any  $i \in [n]$ ,

$$\mu_i = \mathbb{E}_M[f_i] = (\mathbb{E}_M[f'_i] + 1)/2.$$

Recall Eq. (1) that

$$\begin{aligned} \sigma_{ij} &= 2\mathbb{E}_M[f_i f_j] + 1 - \mu_i - \mu_j \\ &= 2\mathbb{E}_M\left[\frac{1 + f'_i}{2} \frac{1 + f'_j}{2}\right] + 1 - \frac{1 + \mathbb{E}_M[f'_i]}{2} - \frac{1 + \mathbb{E}_M[f'_j]}{2} \\ &= \frac{1}{2}\mathbb{E}_M[f'_i f'_j] + \frac{1}{2}. \end{aligned} \quad (10)$$

Thus in the following contents, we focus on calculating  $\mathbb{E}_M[f'_i f'_j]$ . It can be considered by combining the following two cases: a) two coalescing random walkers from  $v_i$  and  $v_j$  walking to the same node in  $V'$ , and b) two coalescing random walkers from  $v_i$  and  $v_j$  walking to different nodes in  $V'$ . Notice that if two random walkers end at the same node  $v'_k \in V'$ , thus  $\mathbb{E}_O[f'_k f'_k] = 1$  for any  $k \in [n]$ ; and if two random walkers end at the different nodes  $v'_k$  and  $v'_l$ , thus

$\mathbb{E}_O[f'_k f'_l] = (2\mu^{(0)} - 1)^2$  for any  $k, l \in [n]$ . Therefore,

$$\begin{aligned} \mathbb{E}_M[f'_i f'_j] &= [p_{\text{same}}^+(i, j) - p_{\text{same}}^-(i, j)] \cdot 1 \\ &\quad + [p_{\text{diff}}^+(i, j) - p_{\text{diff}}^-(i, j)] \cdot (2\mu^{(0)} - 1)^2. \end{aligned} \quad (11)$$

Next, we calculate  $p_{\text{same}}^+(i, j) - p_{\text{same}}^-(i, j)$  and  $p_{\text{diff}}^+(i, j) - p_{\text{diff}}^-(i, j)$  separately, similar to the proof of Lemma 6.

(a)  $p_{\text{same}}^+(i, j) - p_{\text{same}}^-(i, j)$ .

If two walkers starting from  $v_i$  and  $v_j$  walk to the same node in  $V'$ , there are two cases: One is they meet at some node  $v_l \in V$  and then walk together until being absorbed; The other is they do not meet before they end their walks. Thus we have

$$p_{\text{same}}^+(i, j) = \sum_{k=1}^n \sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ij}^{l+}] (Q_{lk}^+ + Q_{lk}^-) + \sum_{k=1}^n \mathbb{P}[\mathcal{H}_{ij}^{k+}],$$

and

$$p_{\text{same}}^-(i, j) = \sum_{k=1}^n \sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ij}^{l-}] (Q_{lk}^+ + Q_{lk}^-) + \sum_{k=1}^n \mathbb{P}[\mathcal{H}_{ij}^{k-}].$$

Consider two non-coalescing random walks starting from  $v_i$  and  $v_j$  are absorbed by the same node  $v'_k \in V'$  while the summation of two walks' steps is even:

$$\begin{aligned} Q_{ik}^+ Q_{jk}^+ + Q_{ik}^- Q_{jk}^- &= \sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ij}^{l+}] (Q_{lk}^{+2} + Q_{lk}^{-2}) \\ &\quad + \sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ij}^{l-}] \cdot 2Q_{lk}^+ Q_{lk}^- + \mathbb{P}[\mathcal{H}_{ij}^{k+}], \end{aligned}$$

Consider two non-coalescing random walks starting from  $v_i$  and  $v_j$  are absorbed by the same node  $v'_k \in V'$  while the

summation of two walks' steps is odd:

$$\begin{aligned} Q_{ik}^+ Q_{jk}^- + Q_{ik}^- Q_{jk}^+ &= \sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ij}^{l-}] (Q_{lk}^{+2} + Q_{lk}^{-2}) \\ &\quad + \sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ij}^{l+}] \cdot 2Q_{lk}^+ Q_{lk}^- + \mathbb{P}[\mathcal{H}_{ij}^{k-}]. \end{aligned}$$

Thus

$$\begin{aligned} p_{same}^+(i, j) - p_{same}^-(i, j) &= \sum_{k=1}^n \sum_{l=1}^n \left( \mathbb{P}[\mathcal{I}_{ij}^{l+}] - \mathbb{P}[\mathcal{I}_{ij}^{l-}] \right) [1 - (Q_{lk}^+ - Q_{lk}^-)^2] \\ &\quad + \sum_{k=1}^n (Q_{ik}^+ - Q_{ik}^-)(Q_{jk}^+ - Q_{jk}^-). \end{aligned} \quad (12)$$

(b)  $p_{diff}^+(i, j) - p_{diff}^-(i, j)$ .

Consider two non-coalescing random walks starting from  $v_i$  and  $v_j$  are absorbed by different nodes in  $V'$  while the summation of two walks' steps is even:

$$\begin{aligned} \sum_{a \neq b} Q_{ia}^+ Q_{jb}^+ + Q_{ia}^- Q_{jb}^- &= \sum_{a \neq b} \sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ij}^{l+}] (Q_{la}^+ Q_{lb}^+ + Q_{la}^- Q_{lb}^-) \\ &\quad + \sum_{a \neq b} \sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ij}^{l-}] (Q_{la}^+ Q_{lb}^- + Q_{la}^- Q_{lb}^+) \\ &\quad + p_{diff}^+(i, j). \end{aligned}$$

Consider two non-coalescing random walks starting from  $v_i$  and  $v_j$  are absorbed by different nodes in  $V'$  while the summation of two walks' steps is odd:

$$\begin{aligned} \sum_{a \neq b} Q_{ia}^+ Q_{jb}^- + Q_{ia}^- Q_{jb}^+ &= \sum_{a \neq b} \sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ij}^{l+}] (Q_{la}^+ Q_{lb}^- + Q_{la}^- Q_{lb}^+) \\ &\quad + \sum_{a \neq b} \sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ij}^{l-}] (Q_{la}^+ Q_{lb}^+ + Q_{la}^- Q_{lb}^-) \\ &\quad + p_{diff}^-(i, j). \end{aligned}$$

Thus

$$\begin{aligned} p_{diff}^+(i, j) - p_{diff}^-(i, j) &= \sum_{a \neq b} (Q_{ia}^+ - Q_{ia}^-)(Q_{jb}^+ - Q_{jb}^-) \\ &\quad - \sum_{a \neq b} \left( \mathbb{P}[\mathcal{I}_{ij}^{l+}] - \mathbb{P}[\mathcal{I}_{ij}^{l-}] \right) (Q_{ia}^+ - Q_{ia}^-)(Q_{jb}^+ - Q_{jb}^-). \end{aligned} \quad (13)$$

In order to obtain  $p_{same}^+(i, j) - p_{same}^-(i, j)$  and  $p_{diff}^+(i, j) - p_{diff}^-(i, j)$ , we need to compute  $\mathbb{P}[\mathcal{I}_{ij}^{l+}]$  and  $\mathbb{P}[\mathcal{I}_{ij}^{l-}]$ . Let  $\mathbb{P}[\mathcal{I}_{ij}^l] = \mathbb{P}[\mathcal{I}_{ij}^{l+}] + \mathbb{P}[\mathcal{I}_{ij}^{l-}]$ ,  $\Delta[\mathcal{I}_{ij}^l] = \mathbb{P}[\mathcal{I}_{ij}^{l+}] - \mathbb{P}[\mathcal{I}_{ij}^{l-}]$ .

Similar to the proof of Lemma 5, we have

$$\begin{aligned} \mathbb{P}[\mathcal{I}_{ij}^{l+}] &= \sum_{a=1}^n \frac{\lambda_i(1-p_i)|A_{ia}|}{2(\lambda_i + \lambda_j)d_i} \left( \mathbb{P}[\mathcal{I}_{aj}^l] + E_{ia}\Delta[\mathcal{I}_{aj}^l] \right) \\ &\quad + \sum_{b=1}^n \frac{\lambda_j(1-p_j)|A_{jb}|}{2(\lambda_i + \lambda_j)d_j} \left( \mathbb{P}[\mathcal{I}_{ib}^l] + E_{jb}\Delta[\mathcal{I}_{ib}^l] \right), \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}[\mathcal{I}_{ij}^{l-}] &= \sum_{a=1}^n \frac{\lambda_i(1-p_i)|A_{ia}|}{2(\lambda_i + \lambda_j)d_i} \left( \mathbb{P}[\mathcal{I}_{aj}^l] - E_{ia}\Delta[\mathcal{I}_{aj}^l] \right) \\ &\quad + \sum_{b=1}^n \frac{\lambda_j(1-p_j)|A_{jb}|}{2(\lambda_i + \lambda_j)d_j} \left( \mathbb{P}[\mathcal{I}_{ib}^l] - E_{jb}\Delta[\mathcal{I}_{ib}^l] \right). \end{aligned}$$

where

$$E_{ij} = \begin{cases} 1, & A_{ij} > 0; \\ -1, & A_{ij} < 0; \\ 0, & A_{ij} = 0. \end{cases}$$

Thus, we can get the recursive equations of  $\mathbb{P}[\mathcal{I}_{ij}^l]$  and  $\Delta[\mathcal{I}_{ij}^l]$ :

$$\begin{aligned} \mathbb{P}[\mathcal{I}_{ij}^l] &= \sum_{a=1}^n \frac{\lambda_i(1-p_i)|A_{ia}|}{(\lambda_i + \lambda_j)d_i} \mathbb{P}[\mathcal{I}_{aj}^l] \\ &\quad + \sum_{b=1}^n \frac{\lambda_j(1-p_j)|A_{jb}|}{(\lambda_i + \lambda_j)d_j} \mathbb{P}[\mathcal{I}_{ib}^l], \end{aligned}$$

and

$$\begin{aligned} \Delta[\mathcal{I}_{ij}^l] &= \sum_{a=1}^n \frac{\lambda_i(1-p_i)|A_{ia}|}{(\lambda_i + \lambda_j)d_i} E_{ia}\Delta[\mathcal{I}_{aj}^l] \\ &\quad + \sum_{b=1}^n \frac{\lambda_j(1-p_j)|A_{jb}|}{(\lambda_i + \lambda_j)d_j} E_{jb}\Delta[\mathcal{I}_{ib}^l], \end{aligned}$$

Then we can get  $\mathbb{P}[\mathcal{I}_{ij}^{l+}] = \frac{\mathbb{P}[\mathcal{I}_{ij}^l] + \Delta[\mathcal{I}_{ij}^l]}{2}$  and  $\mathbb{P}[\mathcal{I}_{ij}^{l-}] = \frac{\mathbb{P}[\mathcal{I}_{ij}^l] - \Delta[\mathcal{I}_{ij}^l]}{2}$  by solving the above two recursive equations.

Above all, we can get opinion similarities between any pair of nodes by Equation (10,11,12,13).